November 2015

Niche-Based Modeling of Japanese Stiltgrass (Microstegium vimineum) Using Presence-Only Information

Nathan Bush
University of Massachusetts Amherst

NICHE-BASED MODELING OF JAPANESE STILTGRASS (*MICROSTEGIUM VIMINEUM*) USING PRESENCE-ONLY INFORMATION

A Thesis Presented

by

NATHAN R. BUSH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
Of the requirements of the degree of

MASTER OF SCIENCE

September 2015

Department of Environmental Conservation

NICHE-BASED MODELING OF JAPANESE STILTGRASS (*MICROSTEGIUM VIMINEUM*) USING PRESENCE-ONLY INFORMATION

A Thesis Presented

by

NATHAN R. BUSH

Approved as to style and content by:

_____
Timothy Randhir, Co-Chair

_____
Charles Schweik, Co-Chair

_____
Cynthia Boettner, Member

_____
Curtice Griffin, Department Head,
Environmental Conservation

DEDICATION

To my dearest parents and loving son.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Timothy Randhir and Dr. Charles Schweik, for their many years of thoughtful and patient guidance and support. I also thank Cynthia Boettner as a committee member, mentor, and colleague. A special thanks to Lena F. and Jason A. whose support and friendship helped me to remain focused. I'd like to thank Charles F., Kate C., and other colleagues for statistical guidance and editing.

ABSTRACT

NICHE-BASED MODELING OF JAPANESE STILTGRASS (*MICROSTEGIUM VIMINEUM*) USING PRESENCE-ONLY INFORMATION

SEPTEMBER 2015

NATHAN R. BUSH

B.S., UNIVERSITY OF MASSACHUSETTS, AMHERST

M.S., UNIVERSITY OF MASSACHUSETTS, AMHERST

Directed by: Professors Timothy Randhir and Charles Schweik

The Connecticut River watershed is experiencing a rapid invasion of aggressive non-native plant species, which threaten watershed function and structure. Volunteer-based monitoring programs such as the University of Massachusetts' OutSmart Invasives Species Project, Early Detection Distribution Mapping System (EDDMapS) and the Invasive Plant Atlas of New England (IPANE) have gathered valuable invasive plant data. These programs provide a unique opportunity for researchers to model invasive plant species utilizing citizen-sourced data. This study took advantage of these large data sources to model invasive plant distribution and to determine environmental and biophysical predictors that are most influential in dispersion, and to identify a suitable presence-only model for use by conservation biologists and land managers at varying spatial scales. This research focused on the invasive plant species of high interest - Japanese stiltgrass (*Mircostegium vimineum*). This was identified as a threat by U.S. Fish and Wildlife Service refuge biologists and refuge managers, but for which no mutli-scale

practical and systematic approach for detection, has yet been developed. Environmental and biophysical variables include factors directly affecting species physiology and locality such as annual temperatures, growing degree days, soil pH, available water supply, elevation, closeness to hydrology and roads, and NDVI. Spatial scales selected for this study include New England (regional), the Connecticut River watershed (watershed), and the U.S. Fish and Wildlife, Silvio O. Conte National Fish and Wildlife Refuge, Salmon River Division (local). At each spatial scale, three software programs were implemented: maximum entropy habitat model by means of the MaxEnt software, ecological niche factor analysis (ENFA) using Openmodeller software, and a generalized linear model (GLM) employed in the statistical software R. Results suggest that each modeling algorithm performance varies among spatial scales. The best fit modeling software designated for each scale will be useful for refuge biologists and managers in determining where to allocate resources and what areas are prone to invasion. Utilizing the regional scale results, managers will understand what areas on a broad-scale are at risk of *M. vimineum* invasion under current climatic variables. The watershed-scale results will be practical for protecting areas designated as most critical for ensuring the persistence of rare and endangered species and their habitats. Furthermore, the local-scale, or fine-scale, analysis will be directly useful for on-the-ground conservation efforts. Managers and biologists can use results to direct resources to areas where *M. vimineum* is most likely to occur to effectively improve early detection rapid response (EDRR).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION AND OBJECTIVES

Introduction

Conservation biologists assert that invasive species may be the greatest threat to current and future biological diversity, ecosystem functions, and the services they provide (Vitousek, D'Antonio, Loope, & Westbrooks, 1996; Mack, Simberloof, Lonsdale, Evans, Cout, & Bazzaz, 2000). Mehroff (2000) estimated that 30-35% of flora in New England is non-native with 3-5% considered non-native invasive species. Many species such as purple loosestrife (*Lythrum salicaria*), and oriental bittersweet (*Celastrus obiculatus*) are common and well known to the general public. However, the more insidious species such as Japanese stiltgrass (*Microstegium vimenium*) and mile-a-minute (*Persicaria perfoliata*) are less known and also less established in northern New England. However, they can potentially cause great damage to vulnerable ecosystems by outcompeting native vegetation, disturbing wildlife habitat, and limiting key resource services. All of these "superplants" not only grow fast and spread quickly, some such as garlic mustard *(Allaria petiolata)* can produce allelochemicals that disrupt and inhibit the growth of ectomycorrhizal fungal communities essential for native plant species growth (Wolfe, Rodgers, Stinson, & Pringle, 2008). Furthermore, some invasive species cause public health problems, such as giant hogweed (*Heracleum mantegazzianum*) which contains phytotoxins within the plant's sap and when exposed to sunlight, cause painful blisters (United States Department of Agriculture, 2012). According to Pimental, Lach, Zungia, & Morrison (2000), public health problems associated with invasive plants cost the United States $36 billion per year.

Microstegium vimineum (Trin.) A. Camus (Japanese stiltgrass) is an annual grass descending from Asia that was first introduced to the United States in Tennessee in 1919 and has become widespread along the eastern United States (Hunt & Zaremba, 1992; Ehrenfeld, 1999). However, it has been yet to be discovered in northeastern states like Vermont, New Hampshire, and Maine. *M. vimineum* has been commonly associated with riparian areas, mesic forests, and disturbed areas such as roads and logging trails (Hunt & Zaremba, 1992). However, Ehrenfeld (1999) describes *M. vimineum* to invade not only wet mesic soils, but also rocky peaks of the Kitatiny Mountains in New Jersey. Due to the lack of the species' cold hardiness and fewer growing degree days in northern New England, *M. vimineum* may be reaching its northern range in the southern states of New England (Hunt & Zaremba, 1992) however, in the lowlands and floodplains of the Connecticut River watershed where warmer temperatures exist, *M. vimineum* may continue to spread northward. Compounded by climate change, regional temperature increases, and microhabitats, this species may cross borders from MA to VT/NH. Considered an invasive colonizer, *M. vimineum* rapidly spreads naturally into these areas via surface storm water flow and animal herbivory, or unintentionally distributed by humans in fill soils, attached to logging equipment, or simply attached to boots and trousers (Gibson & Benedict, 2002; Cole, 2003). Once established, *M. vimineum* can form dense monocultures in forest patches, along streams and roads, and can completely replace native ground cover within 3-5 years (Tennessee Exotic Pest Plant Council , 2013). In locations within New England, such as in parts of the CT River Watershed, and its sub-watersheds, *M. vimineum* is considered an "early detection, rapid response" invasive species because of this destructive potential.

To gain an insight on present and future environmental effects caused by such invasive species, policy-makers rely on the ability to accurately predict the spread and establishment of invasive species- information which is greatly sought after (Ibanez, Silander, Allen, Treanor, & Wilson, 2009). Not only do policy-makers rely on informative models, but refuge managers and biologist will use such information to target areas where invasive species are impacting limited resources and to rapidly respond. This information will also be useful to investigate areas not currently known to have *M. vimineum* present. Given the impact invasive species have upon native species and their habitats, there is an increasing need to develop well-built yet parsimonious models to identify the current extent and predict future spread of invasive plant species. Species distribution modeling (SDM) can identify which areas are prone to invasion and describe biological patterns associated with their physical interactions among local geographic and environmental explanatory variables (Elith & Leathwick, 2009). Of the many SDMs, ecological niche or niche-based modeling (NBM) relies on statistical or theoretical relationships between environmental predictors and observed species distributions. NBM is an appropriate approach when species distribution information is collected as "presence-only" where information is based on species occurrence rather than species "absence", or not occurring. Species data is collected on a binary scale and can be represented in a binomial distribution where the data can be presence, presence/absence, or abundance observations based on random or systematic field sampling. The modeled niche can then be spatial projected or extrapolated into the future using data from general or regional climate models (Morin & Thuiller, 2009).

Niche-based models are useful for organizations such as the Westfield Invasive Species Partnership (WISP) and Cooperative Invasive Species Management Areas (CISMA) to identify high threat geographic areas that are most biologically important and can provide a spatial scheme when deploying strike teams for invasive species control efforts at the watershed and local scales. In 2012 invasive species detection, monitoring, and eradication efforts costs WISP US$ 17,000 (WISP, 2013). The Silvio O. Conte National Fish and Wildlife Refuge, the jurisdictional boundaries of which are delineated by the 7.2 million acre Connecticut River watershed spanning four states within New England, employ these types of models to effectively support the goals and objectives outlined in the National Strategy for the Management of Invasive Species (National Wildlife Refuge System, 2003). According to the U.S. Fish and Wildlife Service (2009), invasive species "cost an estimated $137 billion a year in losses to agriculture, industry, forestry, commercial fishing, recreation, and water supplies" (para. 8). The Department of the Interior spent US$100 million in 2011 for invasive species prevention including EDRR efforts, research, outreach, restoration, and partnership cooperation (United Sates Fish and Wildlife Service, 2012). In 2012, the Silvio O. Conte NFWR spent nearly US$ 20,000 in similar efforts within the Connecticut River watershed for a single species (Boettner C. , 2013) . Likewise, the Connecticut River Watershed Invasive Species Initiative, a partnership between federal, state, local agencies, and nongovernment organizations, that assist with seven sub-watershed-scale Cooperative Invasive Species Management Areas (CISMA) throughout the Connecticut River watershed aims to protect local rare and endangered species and their habitats and to enhance biodiversity. In 2012, the Initiative spent an estimated US$ 50,000 on EDRR,

outreach and other efforts similar to the DOI's prevention measures. Having a precise and accurate model explaining the distribution at multiple scales will highly benefit EDRR and strike team efforts not only at the local scale, but at the larger watershed scale. Lastly, the regional-scale analysis can be used to gain an overall understanding of *M. vimineum* distribution within New England based on current climactic conditions and can be used as a reference distribution for extrapolation and or interpolation under future climactic conditions.

Objectives

The goal of this research was to examine distribution of *M. vimineum* in New England utilizing niche-based modeling techniques and applying environmental and biophysical predictors across multiple landscape scales. Specifically, this study intended to:

1: To study the spatial distribution and suitable habitat of *M. vimineum* at three spatial scales (local, watershed, and regional).

Ha: Distribution and suitable habitat of *M. vimineum* is significantly influenced by scale-specific environmental and biophysical predictors.

2: To develop and implement correlative models for *M. vimineum* predictions at three spatial scales (local, watershed, and regional).

Ha: There exist significant differences between models in the prediction and accuracy of *M. vimineum* occurrence and habitat suitability at each scale.

Significance of Research

This study presents information about how species distribution modeling varies among spatial scales. It also assists in understanding which environmental and biophysical predictors are most relevant when building and implementing the proposed open-source modeling software algorithms at various spatial scales. This research helps fill the gap of the assessment of species distribution modeling, specifically ecological niche modeling, among spatial scales using presence-only citizen-sourced data by testing several commonly used presence-only modeling techniques. The techniques used to assess modeling software in this study can essentially guide refuge biologist and managers with an appropriate method in determining which open-source modeling software and predictor variables should be utilized when coupled with species presence data. The results from this study specify where current *M. vimineum* infestations are likely to occur. Having implemented these parsimonious models, results can tentatively guide managers and refuge staff where to efficiently allocate resources.

II

LITERATURE REVIEW

The following review of literature is broken into three main sections: (1) Issues related to invasive plant species, (2) Current invasive plant species distribution modeling techniques, (3) Current invasive plant management strategies

Issues Related to Invasive Plant Species

The most commonly recognized impact invasive species has on the environment is their ability to suppress native species populations and reduce biodiversity (Wilcove, Rothstein, Dubow, Phillips, & Losos, 1998; Chornesky & Randall, 2003). However, they also can cause impairments or even completely destroy ecosystem functions, services, and integrity by outcompeting native species, disrupting genetic diversity by hybridization, complete invasion of an area, or carry diseases (Council for Agriculture Science and Technology, 2002). Vitousek, (1990) argues that ecosystem-level invaders "alter the fundamental rules of existence for all organisms in the area" (p. 8). Vitousek explains how the plant *Mryica faya*, non-native to Hawai'i Volcanoes National Park, altered fundamental ecosystem-level characteristics by adding a symbiotic nitrogen fixer to a nitrogen limited location, therefore disrupting a primary successional ecosystem (Vitousek, 1990). Pajchar and Mooney piece together the use of mechanistic functions by invasive species, as discussed by the Council for Agriculture Science and Technology, to achieve a competitive edge, furthermore, linking them to the ecosystem services being compromised (Pajchar & Mooney, 2009)

Compromised species populations and ecosystems have huge impacts on the national and even global economy. However, to maintain geographic and research

integrity, I will only discuss impacts of invasive species on the United States economy. Pimentel, Zuniga, & Morrison (2005) found that invasive species cost the United States $120 billion per year, this is conflicting with a previous estimate of $1.1 billion per year by The Office of Technology Assessment (U.S. Congress, Office of Technology Assessment, 1993). However, the 2005 study was calculated based on ten times as many species as the OTA's 79 species study (Pimentel, Zuniga, & Morrison, 2005). This may still be an incomplete estimation because there are nearly 50,000 non-native species in the United States and no single entity is keeping a comprehensive assemblage of costs (Council for Agriculture Science and Technology, 2002; Pimentel, Zuniga, & Morrison, 2005). However, the eight agency member organization, National Invasive Species Council (NISC) established in 1999 has kept detailed records of each of the eight agencies' annual contributions to invasive species activities. These activities include prevention, early detection and rapid response, control and management, research, restoration, education and public awareness, and leadership and international cooperation. The U.S. federal budget for invasive species activities in 2012 was $2.2 billion – an increase of 35% since 2002 (U.S. National Invasive Species Council (NISC), 2013). The United States Department of Agriculture (USDA) contributions total accounts for nearly half of the annual budget each year for each category, and only the Department of Homeland Security contributes more to prevention than any other agency (U.S. National Invasive Species Council (NISC), 2013). The Tennessee company, Invasive Plant Control Inc., is a privately owned business that frequently contracts with the federal, state, and local governments to conduct invasive species consulting and control efforts on federal, state, and municipal lands in Tennessee. The estimated costs per acre of a high

infestation of *M. vimineum* or similar grasses and or forbs range from $219 - $2599 depending on the type of control (chemical or mechanical) implemented (Invasive Plant Control Inc., 2011). These estimates do not include labor costs. With *M. vimineum* creating large dense monocultures, it's easy to identify how control costs alone for a single invasive species can be staggering.

Invasive species can act as a vector for diseases. Vitousek et al., (1996) explain how the Asian tiger mosquito was first introduced into the United States in the 1980's in imported automobile tires for retreading and spread rapidly, infecting 25 states. Feeding on most animals in the United States, the Asian tiger mosquito operates as a vector for the viral infection eastern equine encephalitis, which is commonly fatal to humans (Vitousek, D'Antonio, Loope, & Westbrooks, 1996). Not only do invasive pests carry diseases, invasive plant such as *H. mantegazzianum* or giant hogweed can cause serious human health issues. According to the New York Department of Environmental Conservation (2013), giant hogweed contains photosensitizing furanocoumarins in its sap, which when upon skin contact and exposure to sunlight, may cause a serious skin inflammation called phytophotodermatitis.

Current Invasive Plant Species Distribution Modeling Techniques

Modeling species distributions is enormous research area with many different algorithms and techniques. Some of the earliest methods used environmental envelop models such as Box's 1983 study where he assumed climate is the most significant determinant over biotic interactions for species distributional patterns (Box, 1983). More advanced modeling include machine learning techniques such as multivariate adaptive regression splines which model nonlinearities and interactions between variables in linear

models (Moisen & Frescino, 2002) and artificial neural networks to help identify patterns of associations amongst species (Paini, Bianchi, Northfield, & De Barro, 2011). Although new advancements in species modeling may increase accuracy and precision, implementation of these types of models requires high proficiency in statistical modeling, generally a limited skill for most biologists and land managers, yet in high demand.

Ideally, species distribution data should be collected in a method that includes both presence and absence locations to help model robustness and reduce sampling bias. However, this can be a time consuming and expensive task and some argue that absence locations are misleading because they could potentially indicate detection inability, unsuitable habitat, or suitable habitat that is unoccupied, or false species-environment equilibrium, thus leading to confounding effects (Elith & Leathwick, 2009).

This review section will focus wholly on the use of presence-only data modeling techniques, specifically ecological niche factor analysis, maximum entropy, and generalized linear models coupled with pseudo-absence points. Presence-only modeling consists of utilizing known occurrences of species locations to model species distributional patterns without information of known absences (Elith & Leathwick, 2009). These techniques such as ecological niche factor analysis and BIOCLIM can be categorized as "profile techniques"; however, implementing pseudo-absence points has increased the breadth of models which would generally use true absence points. With pseudo-absence data, "regression-based" models such as GLMs and GAMs and "machine learning" models such as MaxEnt and Random Forest can be used to compare distributional outcomes of accuracy and precision.

Ecological Niche Factor Analysis

Ecological niche factor analysis (ENFA), as described by Hirzel, Hausser, Chessel, & Perrin (2002), is an approach to deal with difficulties surounding absence data as explained by the Elith and Leathwick 2009 review by using presece-only without psuedo-absence data. ENFA, implemented via Biomapper software designed by Hirzel, Hausser, Chessel, & Perrin (2002), compares the species distribution (known precense) within the ecogeographical variable's extent with that of the entire area. This is accomplished by summarizing the overall information into two types of factors, marginality and specialisation. Marginality is the direction in which the species niche differs from the available conditions in the study area. The higher the absolute value of marginality, the more species habitat differs from study area. The specialisation factor indicates how restricted the species' niche is in relation to the study area, i.e., how the species' variance differs from the overall variance of the study area (Ortega-Huerta & Townsend Peterson, 2008; Hirzel, Hausser, Chessel, & Perrin, 2002). Therefore, a maginality factor of one means that the species' habitat is very particular in relation to the "background" or study area, and a high specialization factor indicates a very limited range within the study area. This technique is similair to principal components analysis where a we create a few essential variables, which is a compostion of much of the origianl variables that are most significant in capturing the variation in the dataset and are uncorrelated, thus relieving effects of muticollinearity (Gotelli & Ellison, 2013)

Hirzel, Hausser, Chessel, & Perrin (2002) proposed ENFA as a means to avoid the difficulties associated with absense data. They studied alpine ibex of the Swiss Alpes and found that ENFA predictor computation correlated precisely with the variables found

to be relveant in the known literature for ibex ecology (Hirzel, Hausser, Chessel, & Perrin, 2002). Comparing the ENFA approach to more traditional methods like logistic regression, the team found that their approach did not rely on absence data, which can bias result in GLMs or is logistically difficult to obtain. They also find that traditional methods for variable selection such as stepwise analysis to be sensitive to input order and many trials are needed to extract the "best fit" model. Rather than rejecting variables in traditional stepwise methods, ENFA simply wieghs them for significance (Hirzel, Hausser, Chessel, & Perrin, 2002).

Xuezhi, Weihua, Zhiyun, Jianguo, Yi, & Youping (2008) used ENFA to study the Chinese giant panda habitat selection and associated niche factors. Their results were constistant with an earlier study conducted by the Wanglang Nature Reserve in China, however, some habitat over-estimation may have occurred due to limited bamboo data points (Xuezhi, Weihua, Zhiyun, Jianguo, Yi, Youping, 2008). In a 2013 study, researchers employed ENFA to define habitable locations of Persian leopards, based on 10 uncorrelated environmental factors and presence-only known locations. Researchers found that Persian leopard suitable habitat, defined by the ENFA sutiability model, was in agreement with previous studies of Persian leopard habitat niche (Erfanian, Hamed Mirkarimi, Salman Mahini, & Reza Rezaei, 2013).

When ENFA was tested against Manhalanobis typically, Neeti, Vaclavik, & Niphadkar, (2007) found that Manhalanobis had better ovverall performance in predicting locations of Japanese knotweed in Massachusetts when evaluated by the relative operating characteristics (ROC). In a very similar study, Vaclavik & Ortega (2008) found that ENFA's overall performance was better than Manhalanobis typically

when predicting locations of Norway maple in Massachusetts. Both studies used the same amount of presence points (103 and 104 respectively), and the same amount of environmental predictors. Although 3 of the 8 predictors used in each study were the same, the 5 dissimilar predictors could be more significant to the each species' niche and could explain opposing results. Nonetheless, ENFA appears to be a useful tool when presence-only data is available, and when the variable selection processes such as stepwise analysis is beyond the capabilities of statistically untrained biologists.

Generalized Linear Modeling

Generalized linear modeling (GLM) is a technique that allows for the response variable to be non-normally distributed such as binary data which would form a logistic curve or "S". Because the responese variable in this research is "present" or "absent" or bianry in nature, this review will focus on logistic regression using a logit link function. There are three elements to a GLM. The first element is the response variable and its probability distribution and in this case would be a binomial distribution given that Y represents a binary dataset of "present" or "absent". The second piece is the predictor variables which can be continuioous such as weight or height, or variables can be categorical such as harvesting intesity 1, 2, 3… The third component is the link function which links the response variable and the predictor variable (Quinn & Keough, 2002). Assumptions of logistic regression are met by the binomial prior probabitlity distribution of the response variable, which is likely for bianary data. The logit link of the left side of logistic regression equation:

$$\log\left[\frac{\mu}{1-\mu}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

where $\log\left[\frac{\mu}{1-\mu}\right]$ is the natural log of presence probability over 1 – presence probability and the β_0, β_1, etc., are the parameters to be estimated using maximum likeihood (Gotelli & Ellison, 2013).

In a Canadian study, Weaver, Conway, & Fortin (2012) investigated which environmental predictors that best explain mute swan distribution at several spatial scales reflecting different habitat use and biological activities such as breeding and dispersal. They modeled swan distribution using a GLM with a logit link for binomial dependent variable (presence points and pseudo-absence points) utilizing the statistical software R. The researchers found that the best fit model, acoording to Akaike's information criterion (AIC), utilized all 7 environmental predictors and ranked highest at each scale (Weaver, Conway, & Fortin, 2012).

When GLM was tested among three other models (GAM, Classification Trees, and Random Forests) to examine the importance of explainatory variables influencing the presence or absence of a compilation of 45 plant species in southern California, Syphard and Franklin (2009) found that GLMs, GAMs, and Random Forests had equal prediction accuracy. They also found that for most species, climate variables had higher model importance than topographical or geological variables, suggesting that climate is the main driver for species distribution at large spatial scales (Syphard & Franklin, 2009).
In an effort to model threaten tree species in Morocco, Rupprecht, Oldeland, and Finckh (2011) compared three modeling techniues (GLM, ENFA, and MAXENT) utilizing presence-only data. The three models were evaluated using minimal predicted area (MPA) proposed by Engler, Guisan and Rechsteiner (2004) where a good habitat suitablity map should predict an area that is as small as possible, but includes 90% of

occurance data. Although the results suggested that all three model performed very well, according to the MPA scores, MaxEnt gave the best results considering accuracy and prediction success (Rupprecht, Oldeland, & Finckh, 2011).

Maximum Entropy

Maximum entropy (MaxEnt) has its roots in machine learning, but more recently has been used to model speceis ditributions with presence-only data (Phillips, Anderson, & Schapire, 2006; Blank & Blaustien, 2012). MaxEnt estimates the probability distribution from incomplete data and operates of a set of constraints from what is known from the training data. MaxEnt diferentiates the background environment or areas of possible absence, with a set of points. These points however, could populate the same space as an unknown presence points (Blank & Blaustien, 2012). The algorithm predicts the probibilty distribution across the entire study area and implements maximum entropy principles and regularization parameters to prevent over-fitting (Phillips, Anderson, & Schapire, 2006). More detailed information about maximum entropy principles and initial testing and for in-depth explanation of MaxEnt in species distribution modeling are described in Phillips et al. (2006) and Elith et al. (2011).

Predictions are most often reported as relative logistic probabilities ranging from 0 to 1. The validation of model outputs from MaxEnt is accomplished by defining a percentage of the data for model testing and plots testing and training omissions against an AUC threshold. Finally, MaxEnt will generate response curves for each predictor variables.

Blank and Blaustien (2012) utilize MaxEnt software version 3.3.3e developed by Phillips et al. (2004) to model endangered amphibians in Isreal using limited presence-

only data. They found that even though a very small sample size was used, coupled with local environmental predictors, MaxEnt provides precise and accurate species range effectively influencing management decisions and conservation efforts in the region. Phillips, Anderson, & Schapire (2006) implemented MaxEnt and genetic algorithm for rule-set prediction (GARP, also a machine learning, presence-only method) in a continental-wide study of two Neotropical mammals. Using the same environmental predictors, they found that MaxEnt had consistently better AUC scores than GARP and is more useful for producing fine-scale predictions. The higher AUC scores descriminate between suitable and unsiutable areas for the species.

Another study by Kumar & Stohlgren (2009), found that MaxEnt had a 91% success rate and was statistally significant in detecting areas of a threatened tree species. Only 11 presence-only records and a small combination of climate and topographical predictors were used to build the model in MaxEnt. Although the habitat suitablitiy map may be overfitting the potential distribution, this is the first time a threatened tree species in New Caledonia has been modeled, providing higly effective and timely information for managers to make educated decisions (Kumar & Stohlgren, 2009).

Current Invasive Plant Management Strategies

The Chief of the USDA Forest Service has deemed invasive species as one of the greatest threats to National Forests and rangeland ecosystems (United States Department of Agriculture, Forest Service, 2004). The USDA (2004) has implemented, through the Forest Service, a National Strategy and Implementation Plan for Invasive Species Management that is designed to "reduce, minimize, or eliminate the potential for introduction, establishment, spread, and impact of invasive species across all landscapes

and ownerships". (p. i). The core of the plan is based on four elements: 1) prevention, 2) early detection and rapid response, 3) control and management, 4) rehabilitation and restoration (United States Department of Agriculture, Forest Service, 2004). Similar to the Forest Service's plan, The Department of the Interior's United States Fish and Wildlife Service has a National strategic plan for invasives. The National Strategy for Management of Invasive Species (2003) aims to, "[t]hrough partnerships, prevent, eliminate, or significantly reduce populations of aquatic and terrestrial invasive species throughout the Refuge System in order to protect, restore, and enhance native fish and wildlife species and associated healthy ecosystems." (p. 3). The strategy relies on four main goals: 1) increase the awareness of the invasive species issue internally and externally, 2) reduce impacts of invasive species to allow the Refuge System to more effectively meet its fish and wildlife conservation mission and purpose, 3) reduce impacts of invasive species on Refuge System neighbors and communities, 4) Promote and support the development and use of safe and effective integrated management techniques to combat invasive species (National Wildlife Refuge System, 2003). With strategic plans such as these, subordinate agencies like the United State Geological Survey (USGS) support the Department of the Interior (DOI) and USDA with the research, planning, and management decisions. It's mission statement, according to the USGS (2004), is "[t]o provide reliable information and useful tools for documenting, understanding, predicting, assessing, and addressing threats from invasive species in U.S. ecosystems." (p. 10).

Following the guidelines of the National Strategy for Management of Invasive Species, the Silvio O. Conte National Fish and Wildlife Refuge employs a full-time

Invasive Plant Control Initiative Coordinator who's responsible for outreach, education, and building partnerships with local federal, state, and NGOs within the Connecticut River watershed (Silvio O. Conte National Fish and Wildlife Refuge, 2014). Invasive species control efforts within the Silvio O.Conte NFWR include the multi-partner water chestnut (*Trapa natans*) project, where seasonal staff and volunteers locate and hand-pull water chestnut from water bodies within the Connecticut River watershed. A goal is to develop a protocol for effective early detection and rapid response (EDRR) to other new invaders, but thus far detection efforts have lacked an organized or systematic approach. Thus, new invaders like *M. vimineum* are likely spreading unnoticed.

To make control efforts more efficient and effective, many organizations are building prioritization models to increase efficiency with the current trend of decreased funding. The Connecticut River Invasive Species Partnership has developed a watershed-wide GIS-based analysis of priority areas for invasive species eradication. The analysis uses state-level GIS layers such as areas of high ecological integrity, wetlands, floodplain, and other areas of ecological importance, including analyses of resiliency to climate change (Connecticut River Invasive Species Partnership, 2014). This report will help guide CISMAs to identify areas of local importance within their area. Furthermore, WISP, a local CISMA of the Westfield River watershed, has conducted sub-watershed scale GIS-based prioritization analysis to target limited volunteer labor and funding toward the most appropriate and important eradication efforts within its watershed boundaries.

Another proposal that has been gaining traction over the last few years is invasive species "strike teams". The proposal is based on a number of such teams operating across

the country. One example is the New Jersey Strike Team whose mission is to prevent the spread of emerging invasive species by engaging public and private land stewards to implement EDRR tactics (NJ Invasive Species Strike Team, 2014). Applying this model to the Connecticut River watershed, which spans large areas in four states and thus needs to take into consideration different state regulations and partners, proves to be a daunting task. One such partner, Dr. Charles Schweik of the University of Massachusetts, Amherst, has proposed to utilize partner colleges within the watershed. Undergraduate students interested in invasive species would act as the strike team under the supervision of a funded graduate student and a local state or federal employee. Having multiple partnering colleges within the watershed, "strike teams" could be deployed more readily when an outbreak of an emerging invasive is reported.

In order to better predict where invasive species are likely to exist and thus more effectively deploy strike teams, many researchers are coming up with intuitive methods. Species distribution models prove to be a valuable tool, but are only useful to those with statistical background. Open-source software that is readily accessible to the public and models that do not require a great deal of statistical knowledge is more likely to be employed by on-the-ground organizations. The open-source software and parsimonious model described above have the potential to greatly increase EDRR efficiency by identifying ares most likely to harbor the target species.

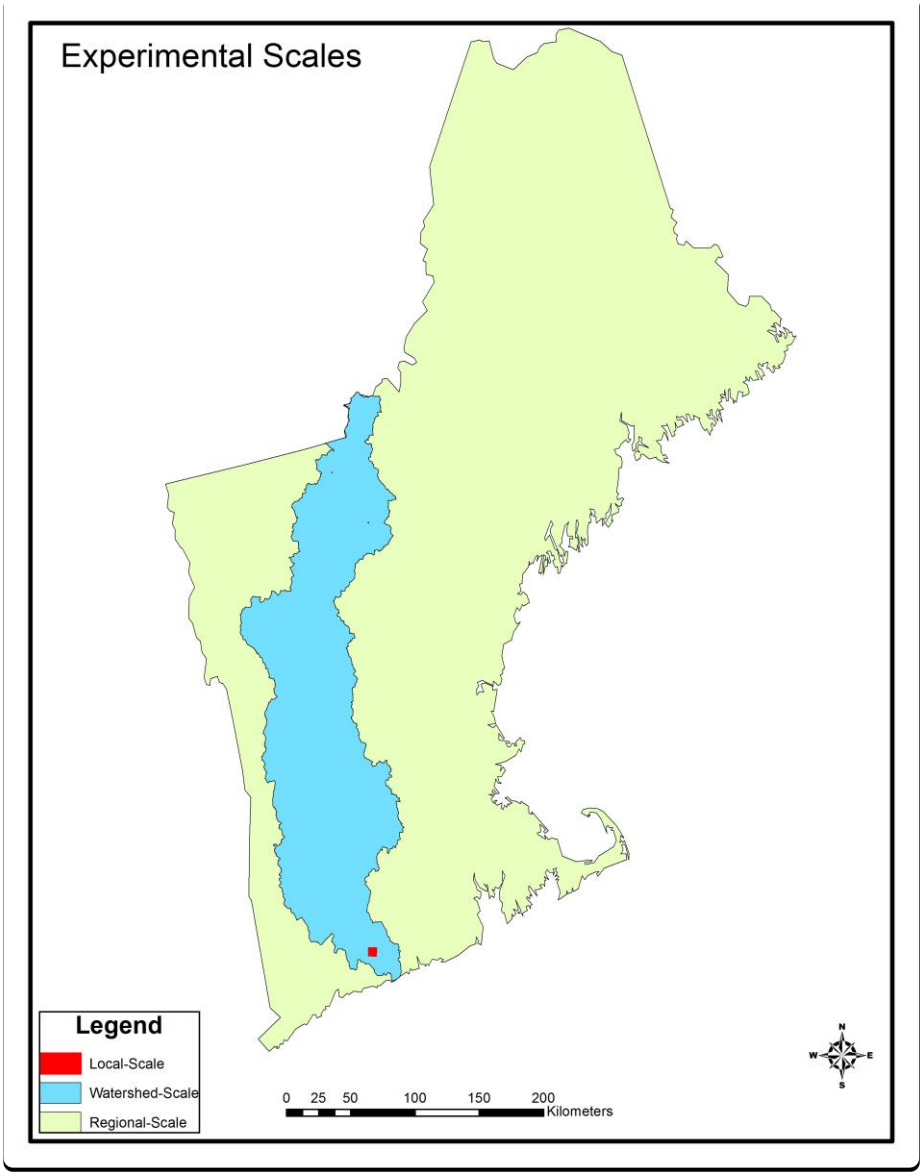
III

METHODOLOGY

Study Site

This study is a multi-scale analysis of the probability of occurrence and habitat suitability of *M. vimineum* influenced by scale-specific climate, topographic, environmental and biophysical predictors. The study took place in the northeastern portion of the United States (Figure 1). Predictors of invasive species presence were examined at the regional, watershed, and local scale. The regional scale includes the six states that comprise the New England region (Connecticut, Massachusetts, Rhode Island, Maine, New Hampshire, and Vermont). States were not analyzed individually, but rather as part of the regional and or watershed scale. The watershed-scale encompasses the boundaries of the Connecticut River watershed, covering a large portion of VT, NH, MA, and CT. This watershed has a hydrologic unit code (HUC) of 8 that covers more than 2.9 million square hectares. Lastly, the local scale focused on the Silvio O. Conte NFWR Salmon River Division located in lower central Connecticut. The Salmon River Division contains nearly 3000 acres of ecologically significant wildlife habitat within the overall Silvio O. Conte NFWR.

Figure 1: Map displaying the boundaries of the three scales of the study area: Regional (New England), Watershed (Connecticut River watershed), local (Salmon River Division).



Presence-Only Data Collection

Occurrence data for *M. vimineum*, in the form of known presence locations, was extracted from the Early Detection and Distribution Mapping System (EDDMapS)

(University of Georgia, Center for Invasive Species and Ecosystem Health, 2014). EDDMapS, launched in 2005 by the Center for Invasive Species and Ecosystem Health at the University of Georgia, is a web-based mapping system for reporting invasive species throughout the nation. Currently, EDDMapS has nearly 2.2 million records of invasive species nation-wide. This data is derived from a combination of organizations and agencies, and volunteers to form a freely accessible species distribution data for interested researchers, educators, land managers, and biologists (University of Georgia, Center for Invasive Species and Ecosystem Health, 2014) EDDMapS incorporates data from multiple sources such as Invasive Plant Atlas of New England (IPANE) (University of Georgia, Center for Invasive Species and Ecosystem Health, 2014) and the Outsmart Invasive Species Project (University of Massachusetts, 2015). IPANE is similar to EDDMaPS in that the data is mostly volunteer-obtained, but IPANE was developed to be a web-accessible database for invasive plants specifically within New England (University of Georgia, Center for Invasive Species and Ecosystem Health, 2014). The Outsmart Invasive Species Project is a partnership between the University of Massachusetts Amherst, the Massachusetts Department of Conservation and Recreation (MA DCR) and the Center for Invasive Species and Ecosystem Health at the University of Georgia. The goal of the project is to strengthen ongoing invasive-species monitoring efforts in New England by utilizing crowd-sourcing technology. This web- and mobile app-based approach enables users to identify species via text and images, or high quality embedded instructional videos, and to make reports to the national database EDDMapS directly from any portable smartphone or tablet utilizing the device's internal GPS capabilities.

Predictor Variables

Predictor variables were collected from online federal, state, and educational institution geographic information system (GIS) departments such as U.S. Geological Survey (United States Geological Survey, 2014), Massachusetts Office of Geographic Information (Office of Geographic information, 2015), and the University of Massachusetts, Amherst (University of Massachusetts, 2000). Variables were grouped into three main categories that are likely to explain *M. vimineum* distribution at each of the three scales: 1) climate variables, 2) topographic and landscape variables, 3) local and fine-scale predictors.

Climate variables were extracted from the Oregon State University's PRISM Climate Group (Northwest Alliance for Computational Science and Engineering, 2015). These include minimum, maximum, and mean temperature yearly averages over a 30 year period (1981-2010). Growing degree days were obtained through the University of Massachusetts, Amherst, Landscape Ecology Lab's Designing Sustainable Landscapes project (University of Massachusetts, 2000). Growing degree days represent the number of days in which the average temperature is above 10 degrees Celsius. This dataset was projected for 2010 by using the PRISM 30 year climate data. All datasets were reprojected to a 100 meter resolution using the "project raster" tool with bilinear interpolation in ArcGIS 10.2 (ESRI, 2013).

Topographic and landscape variables were collected from the University of Massachusetts, Amherst, Landscape Ecology Lab's Designing Sustainable Landscapes project, the United States Geologic Service's National Elevation and hydrography

Datasets data portals, and derivatives thereof. Topographic and landscape variables were collected at a 30 meter resolution. These include:

(* indicates datasets used to derive other variables)

Designing Sustainable Landscapes' datasets

- Topographical wetness: Soil moisture, measured by a topographic wetness index, and based off the Freeman FD8 flow accumulation model.
- Soil available water supply: The total volume of water (cm) that is available to plants in the soil. Calculated as the available water capacity, times the thickness of each horizon to a specified depth of 25 cm. Derived from Natural Resource Conservation Service's STATSGO2.
- Incident solar radiation: Based on a custom algorithm utilizing geographic location, slope, aspect, and topographic shading.
- Soil pH: Measures acidity. Derived from Natural Resource Conservation Service's STATSGO2.
- *Hard development: Includes impervious surfaces such as roads, trains, barren land, and high intensity development. Derived from The Nature Conservancy's Ecological Systems Model Plus

USGS datasets

- *Elevation: 1 arc-second (30 meters). Derived from LiDAR projects.
- *Hydrography: Represent surface water such as rivers, streams, canals, lakes, ponds, coastlines, dams, and stream gauges. Derived from the elevation dataset (1:24000-scale)

Derived datasets

- Aspect: Using the USGS elevation dataset, the ArcGIS 10.2 "aspect" tool creates a raster surface of slope direction with values in the compass direction (0-360 degrees).
- Distance to hydrologic features: Utilizing the "Euclidean distance" tool in ArcGIS 10.2 to calculate an index of distance from water features based on a combination of the USGS hydrography dataset and the United States Fish and Wildlife Service's National Wetland Inventory.
- Distance to hard feature: Using the "Euclidean distance" tool in ArcGIS 10.2 to calculate and index of distance from hard features based on the Designing Sustainable Landscape's hard development dataset.

Local and fine-scale variables were collected at the one meter resolution. Since very few datasets exist at such a fine scale over such a large area, the only freely available dataset was the United States Department of Agriculture's National Agriculture Imagery Program (NAIP) (United States Department of Agriculture, Farm Service Agency, 2015) aerial imagery. The NAIP dataset is a one meter resolution 4-band aerial imagery and was collected in 2014. Each band represents a color bandwidth (band 1 = red, band 2 = green, band 3 = blue, band 4 = near infrared). A normalized difference vegetation index (NDVI) was performed based off the Red and near infrared (NIR) bands. Since chlorophyll highly reflects incoming solar radiation in the near infrared light spectrum, and strongly absorbs light in the normal visible range, this difference can be exploited. NDVI was performed using raster calculator in ArcGIS 10.2 producing a raster with values ranging from -1 to 1 where; $NDVI = (NIR - band\ 1) \div (NIR + band\ 1)$

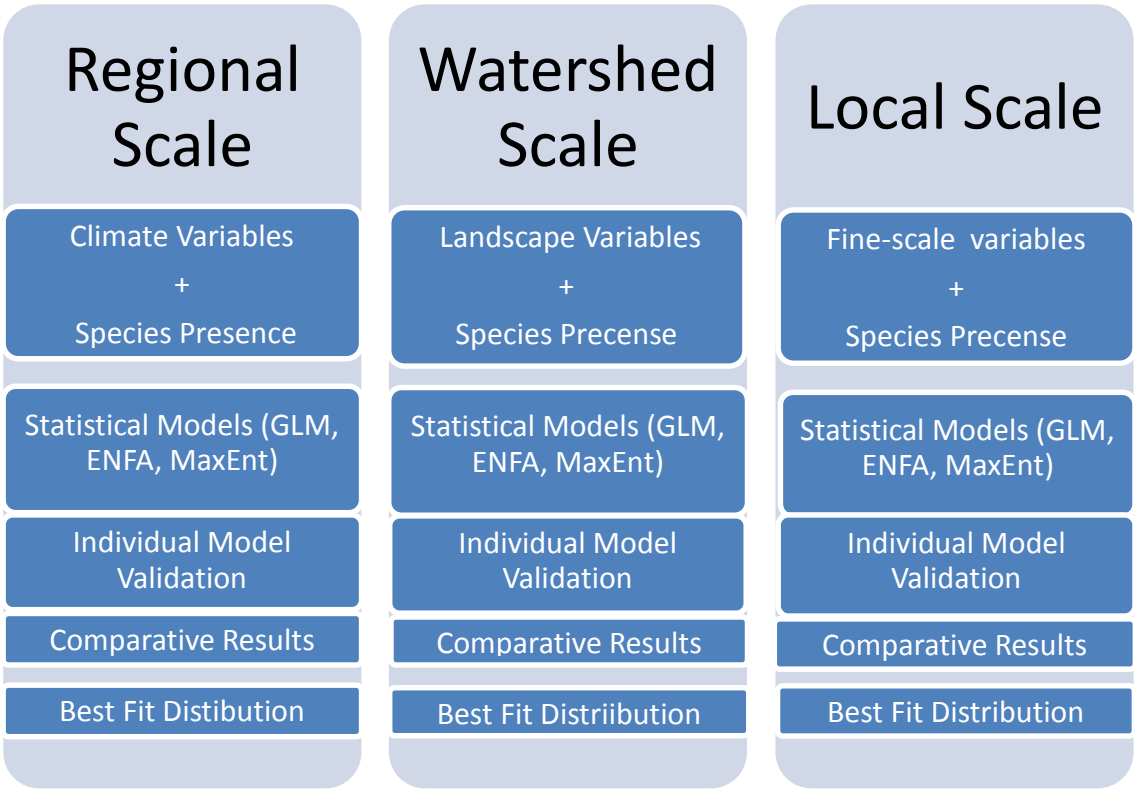
Table 1: Description of predictor variables. Spatial reference indicates spatial scale (regional, watershed, local) (* indicates datasets used to derive other variables).

Description	Spatial Reference	Units
Annual temperature max	Regional	℃
Annual temperature min	Regional	℃
Annual temperature mean	Regional	℃
Annual precipitation	Regional	mm
Growing degree days	Regional	#days
Aspect	Watershed	°NSEW
∗ Distance to hard features	Watershed	m
∗ Distance to water features	Watershed	m
Elevation	Watershed	m
Soil pH	Watershed	#pH
Solar radiance	Watershed	#index
Topographic wetness	Watershed	#index
Soil available water supply	Watershed	#index
*NDVI	Local	#index

Modeling

The modeling was divided into three scales: regional, watershed, and local. At each scale three algorithms (GLM, ENFA, MaxEnt) utilizing presence points were employed with relevant predictors with respect to scale, i.e., climate variables as predictors for the regional-scale and topographic/landscape variables for the watershed-scale. Variable selection and model validation is described in detail under each modeling method.

Figure 2: Conceptual model representation of the modeling and validation process.

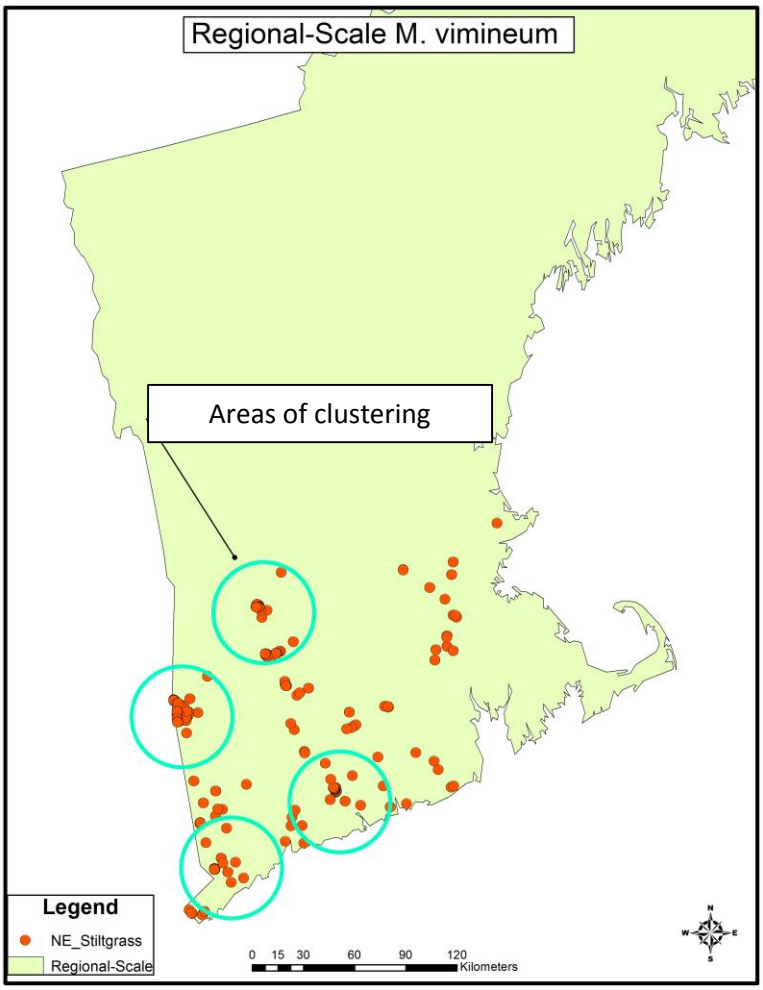


Regional-Scale

Sample Point Evaluation

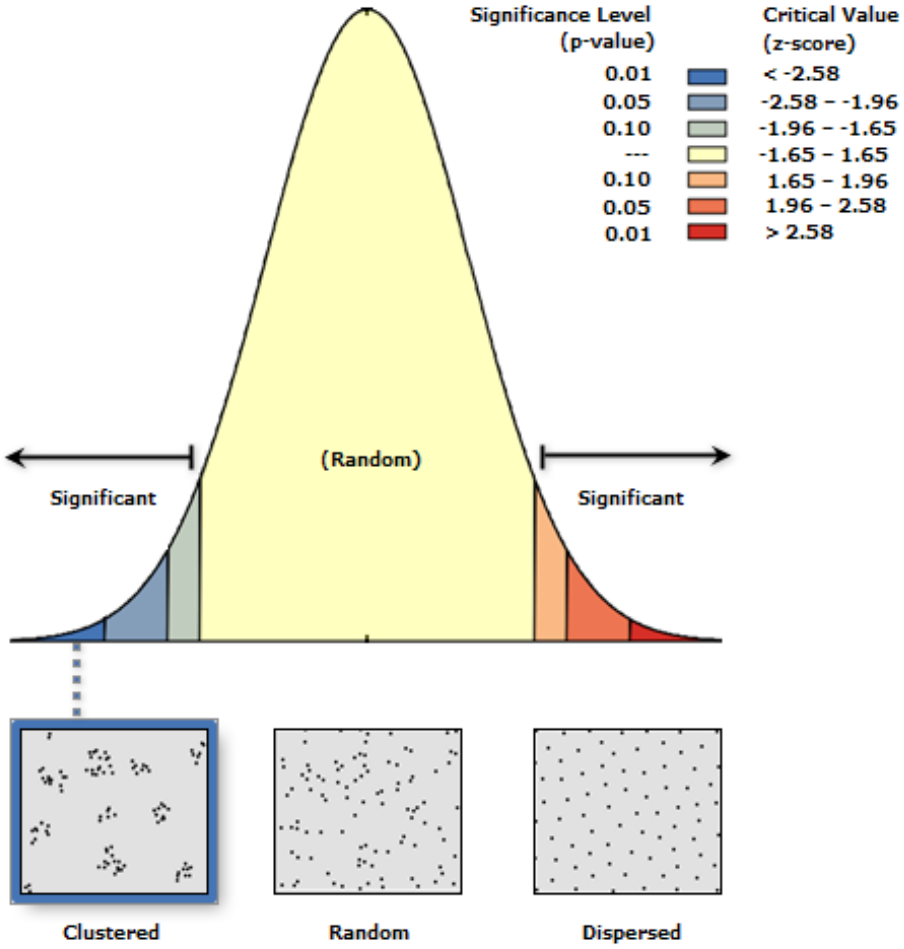
1078 *M. vimineum* points were downloaded from EDDMapS and clipped in GIS to the New England boundary. Visually inspecting the points in GIS, it was clear there were areas of high clustering of points in easily accessible areas such as roads as seen in Figure 3.

Figure 3: Map displaying possible spatial autocorrelation at the regional scale.



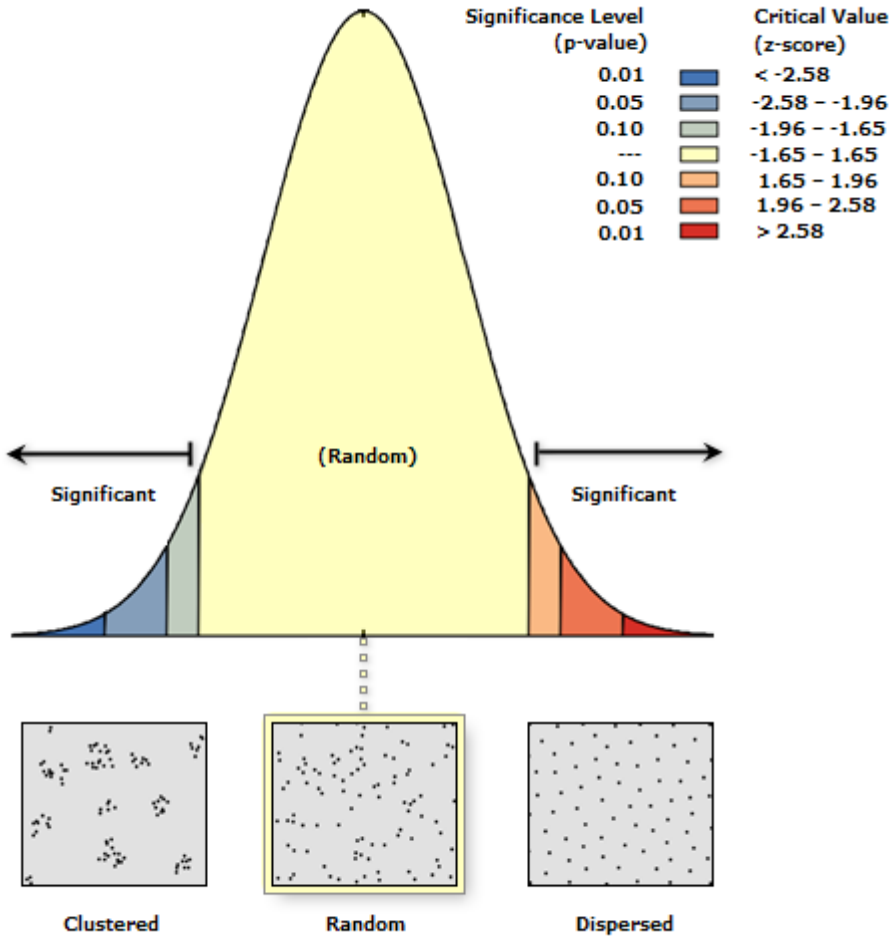
To reduce the adherent clustering, sample points were evaluated in the "Average Nearest Neighbor" tool in ArcGIS 10.2. Results from the nearest neighbor tool suggests the points were significantly clustered with a P-value = 0.0000001, Z-score -44.802 as seen in Figure 4,

Figure 4: Average nearest neighbor output before cluster correction.



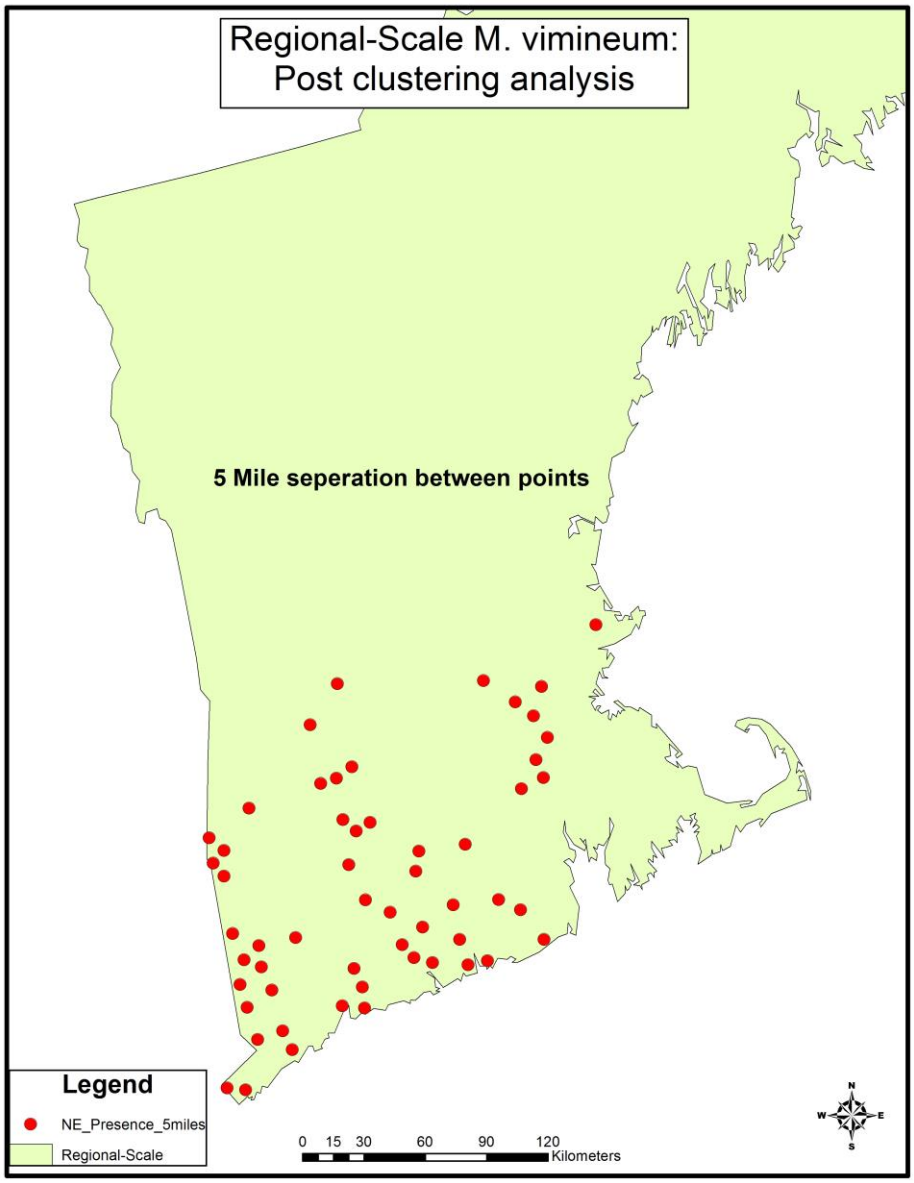
To avoid clustering, a point resample with a threshold of 5 miles was conducted in ArcGIS 10.2. This allowed for a non-significant P-value of 0.297527 or no significant indication of clustering and z-score of -1.042 as seen in Figure 5.

Figure 5: Average nearest neighbor output after cluster correction.



Although clustering was eliminated, the sample size left for experimentation was reduced to 56 presence points as seen in Figure 6. The remaining points were used in the validation of the models.

Figure 6: Map displaying post-cluster analysis with a 5 mile separation between points.



Variable Selection

Climate variables were evaluated in the statistical program R (R Core Team, 2014) using the biostats package designed by Dr. Kevin McGarigal of the University of Massachusetts, Amherst (McGarigal, 2013). A scatter plot matrix (SPLOM) with the Pearson's correlation coefficient (Figure 7) displays the non-clustered points and the variable's direct, indirect, or no evidence of correlation with one another. A cut-off value

of 0.30 (positive or negative) was established to determine correlation. As a result, temperature mean and growing degree days were removed from further analysis as having the highest correlated values among all predictors. Five models were constructed from the remaining variables grounded on basic understanding on plant biology.

Model 1 (NEM1): Annual temperature minimum, annual temperature maximum, and annual precipitation.

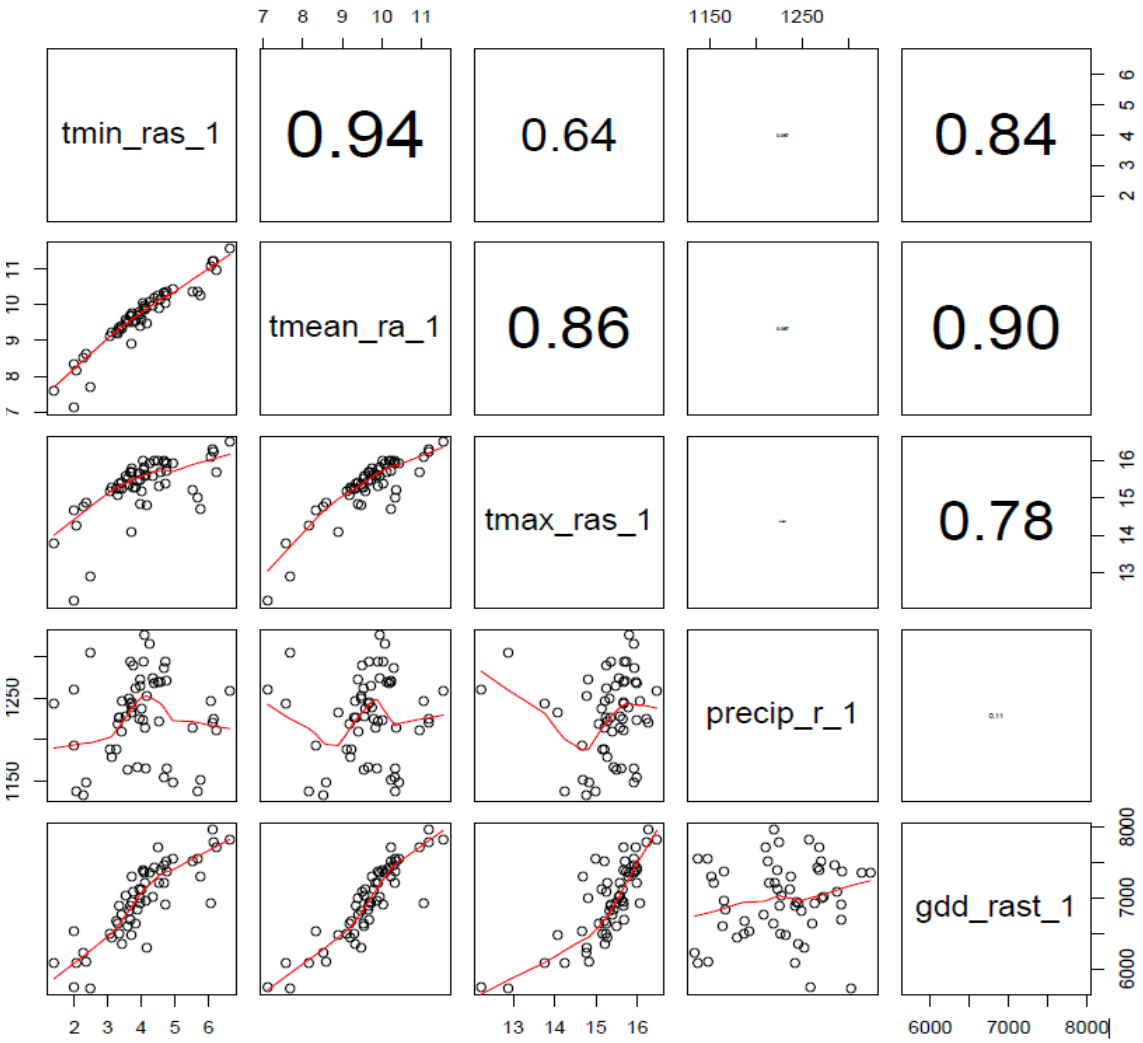
Model 2 (NEM2): Annual temperature maximum, and annual precipitation.

Model 3 (NEM3): Annual temperature minimum, and annual precipitation.

Model 4 (NEM4): Annual temperature minimum, and annual temperature maximum.

Model 5 (NEM5): Annual temperature maximum.

Figure 7: Scatter plot matrix of climate variables and Pearson's correlation coefficient.



Models

Generalized Linear Model

The GLMs were fitted in R software version 2.15.1. Since GLMs require absence points, Berbet-Massin, Jiguet, Albert, & Thuiller (2012) suggests ten thousand randomly generated or pseudo-absence points within the study area will help differentiate where species can and cannot occur. Pseudo-absence points were randomly generated in ArcMap 10.2 using the "create random points" tool with a minimum distance of 1 meter between each point and restricted any random points to fall on actual presence points.

All presence and pseudo-absence points were "merged" together and the predictor cell values where points existed were extracted using the "extract multiple values to points" tool in ArcGIS 10.2. The attribute table was then exported as a .CSV file compatible with R.

Ecological Niche Factor Analysis

Ecological niche factor analysis (ENFA), unlike GLM, does not require pseudo-absence points, yet call for presence-only points and a set of GIS predictor variables. ENFA uses factor analysis to account for multicollinearity among variables, a method used to combine highly correlated observed variables into a single or a few essential unobserved factors or components that represent the linear combination of the observed correlated variables (StatSoft, 2014) . The concept of the marginality and specialization factors as highlighted above and the extraction processes of these factors are described in more detail in Hirzel et al. (2002). ENFA produces habitat suitability index maps which were created from the factors that express the highest percent of variance in the distribution of *M. vimineum.* Habitat suitability maps are derived from a habitat suitability index scaled from 0-100. Models of each scale with their respective predictor group are evaluated by the area under the curve (AUC) of a receiver operating characteristic (ROC) as suggested by Phillips et al. (2006) which illustrates the performance of a binary classification with a threshold and plots the fraction of presence versus absence, in this case randomly drawn background data.

The ENFA models were executed in the freely available open modeling software OpenModeller (Munoz, et al., 2011). The default parameters were accepted and same groups of variables and 56 presence points in each of the five GLM models were applied to the five ENFA models.

Maximum Entropy Algorithm

MaxEnt is a general-purpose machine-learning method that makes predictions and inferences with incomplete data, i.e., presence-only information (Phillips, Anderson, & Schapire, 2006). Given a set of presence-only points, MaxEnt targets a probability distribution by finding the probability distribution that is of maximum entropy, or that is closest to uniform. MaxEnt also samples 10,000 pixels from the study area which are used in the calibration to describe the "background" of niches available to the target species, in this case *M. vimineum* (Anderson & Gonzalez Jr., 2011). The background data informs the model about the density of the predictor variables within the study area allowing for comparison with the density of predictor variables of those occupied by the presence points (Elith, et al., 2011). MaxEnt prevents over-fitting by employing maximum entropy principles and regularization parameters. Further mathematical explanation, and use in species modeling are described in detail in Phillips, Dudik, & Schapire, 2004.

MaxEnt produces probability of suitable habitat in the form of species habitat suitability maps derived from a logistic output ranging from 0 to 1 for each pixel in the study area (Rupprecht, Oldeland, & Finckh, 2011).

Like the GLM and ENFA models, the 5 groupings of variables and 56 presence points were applied to the MaxEnt models. Each of the five model's parameters was adjusted to allow for validation, replication, and optimization. 30 percent of the presence localities were set aside in a random seed method to be used for model validation. The number of replicates was increased from 1 to 15 to allow for more averaging across model runs. Replicated run type was set to the bootstrap method of sample replacement.

The training iterations were increased from 500 to 5000 for more optimization. The default was accepted for all other parameters.

Results

Results from the GLM suggest that model 5 (NEM5) annual temperature maximum was a highly significant predictor (Table 2), the model also had the lowest Akaike information criterion (AIC) score of 669.13 (Table 3). Since the deltaAIC scores were relatively low among most of the models, a weighted model averaging was conducted (Table 4). We can see that model 1 contributed 50% of the explained variance. However, when annual temperature maximum is combined with annual precipitation in model 2, contribution lowered by 22%

To spatial display the results from the GLM models, each model's formula were scripted into the "Raster calculator" tool in ArcGIS 10.2. The resulting rasters were scaled from 0 to1 displaying actual probability of occurrence of *M. vimineum*.

Table 2: GLM model 5 (NEM5) outputs (formula = Abundance ~ tmax, family = binomial, data = nepa).

	Estimate	Std. Error	Z Value	Pr(>|z|)
Intercept	-20.4459	1.7940	-11.397	$< 2e-16$ ***
Temp Maximum	1.11	0.1184	9.398	$< 2e-16$ ***
AIC: 669.1				

Table 3: Akaike information criterion (AIC) – all models.

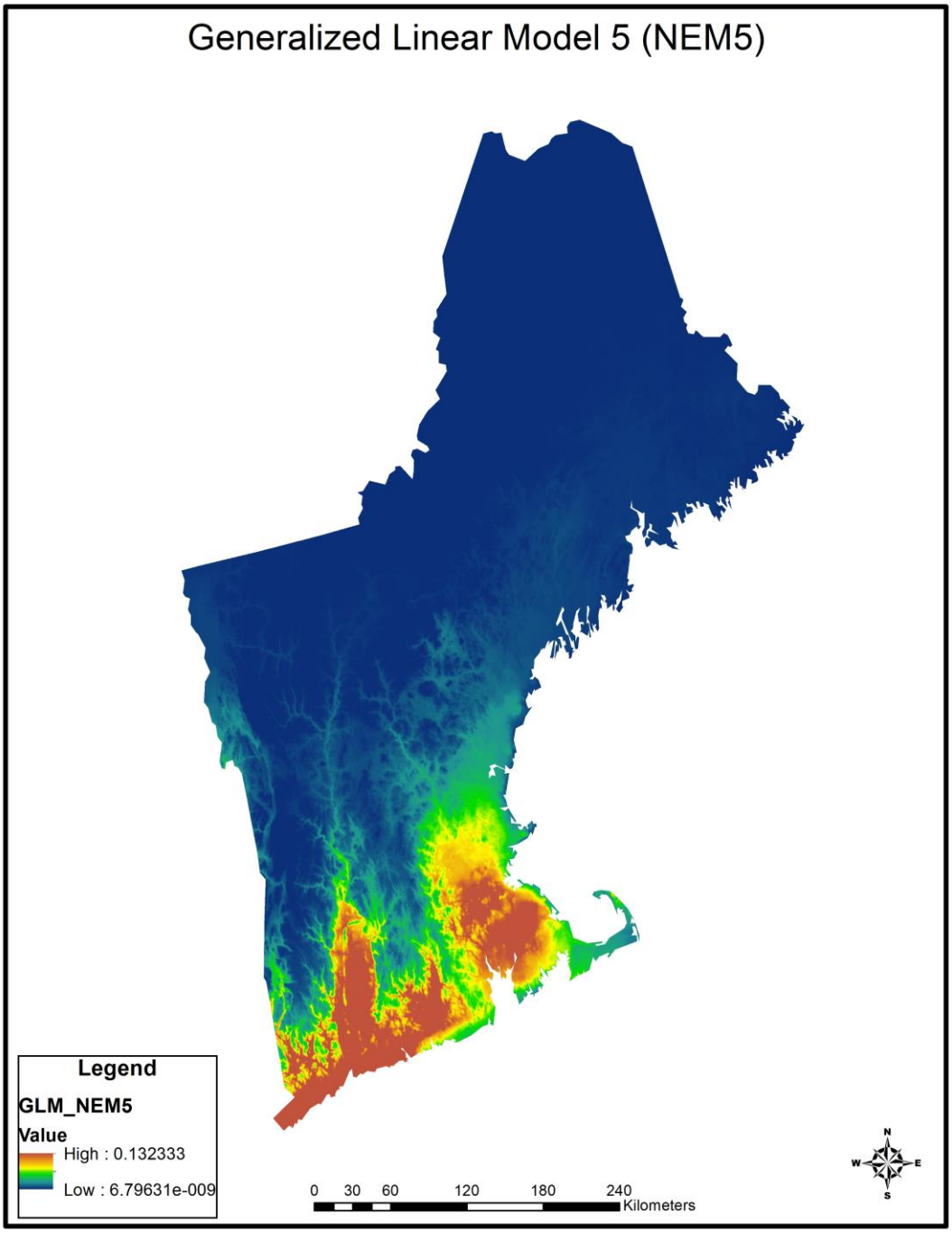
Model	DF	AIC
NEM5	2	669.13
NEM2	3	670.63
NEM4	3	671.17
NEM1	4	672.69
NEM3	3	711.59

Table 4: Model averaging components.

Variables	DF	LogLik	AIC	Delta	Weight
2	2	−332.55	669.11	0.00	0.50
12	3	−332.32	670.65	1.54	0.23
23	3	−332.55	671.10	2.00	0.18
123	4	−332.32	672.65	3.54	0.09
13	3	−352.72	711.44	42.33	0.00

TERM CODES:	Precip	Temp Max	Temp Min
	1	2	3

Figure 8: Results from GLM model 5 (NEM5).



Results of the ENFA based off the AUC score of each model suggest that model 5 (NEM5) performed the best with the highest AUC score of 0.84. Each of the ENFA

model outputs produces two factor values, marginality and specialization. Marginality, as defined by Hirzel et al. (2002), is the difference between the global mean of each component and the species mean within each component, divided by 1.96 of the global mean of each component's distribution. When the marginality factor is close to one, it's suggested that the species lives in a very specialized or extreme habitat relative to the overall conditions.

Hirzel (2002) describes specialization as the ratio of the standard deviation of the component's global distribution to that of the standard deviation of species mean distribution. Any specialization value exceeding one indicates some form of specialization or a specialized niche in comparison to the component.

The broken stick discard method refers to statistician MacArthur's broken stick method for model component retention where a unit of length represents some species' component. The stick is broken into pieces of random lengths, whose resulting length are proportional to the species presence. Components whose value is larger than what would have been obtain at random, or in this case one, are considered significant.
Lastly, factor weight is simply the amount of variation explained by the individual factors.

Models 1 (NEM1), 2 (NEM2), and 3 (NEM3) have all appear to fail and are not evaluated as significant models. Although the models produced marginality and specialization factors, the AUC score were 0.50, or no better than random. This could be a result of the failure of the broken stick discard method due to sampling not covering a large enough range of component values, but a definite cause for model failure was not achieved.

Model 4 (NEM4) originally had two predictor values (annual temperature minimum and annual temperature maximum). The ENFA combines the two predictors into a single uncorrelated component which explain 0.81 of the variance and produced an AUC score of 0.78. Model 5 (NEM5), on the other hand, produced the best AUC score of 0.84 with only a single variable (annual temperature maximum). With a marginality factor 0.75 and specialization factor of 2.15, model 5 (NEM5) suggest that *M. vimineu's* distribution is moderately specialized within the study area.

Figure 9 : ENFA Model 5 (NEM5) area under the curve of the reveiver operating characteristic

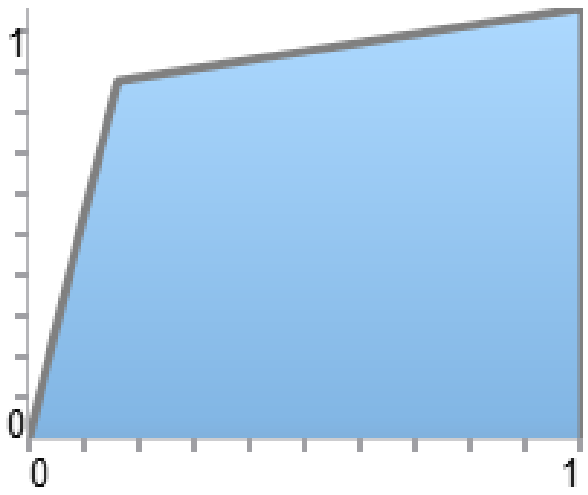
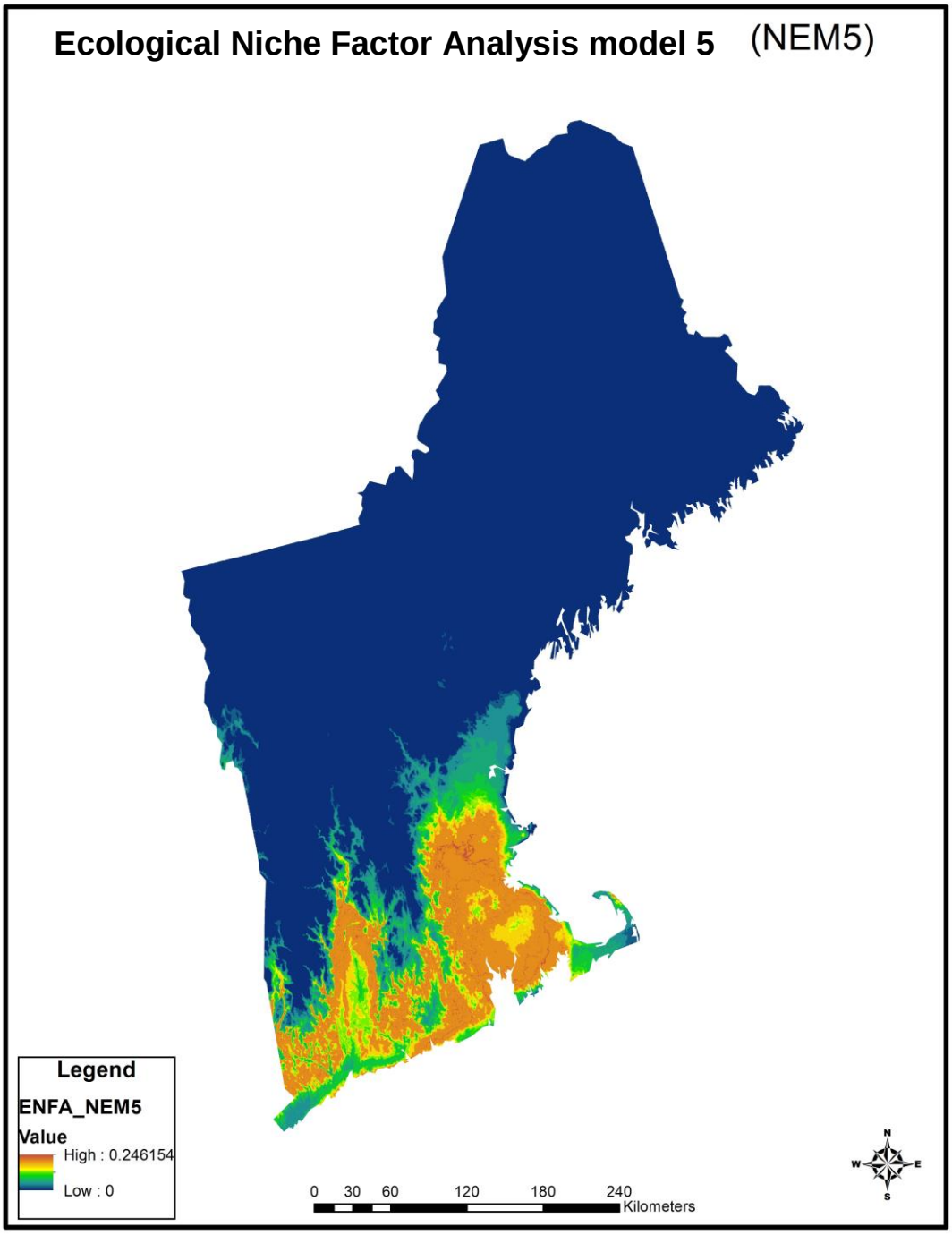


Figure 10: ENFA Model 5 (NEM5) raster habitat suitability map



MaxEnt outputs include area under the curve (AUC) scores which is a measurement of the true positive rate (sensitivity), actual presence points, against the false-positive rate (1-true negative rate or specificity), actual absence points – in this case

pseudo-absence points. A model with an AUC score closer to one indicates better model predictability, those that fall near or below the threshold of 0.5 indicates that a model will perform worse or no better than a random prediction.

MaxEnt also keeps track in the amount in which predictor variables are contributing to the model and their permutation importance. Permutation importance depends on the final MaxEnt model and according to Kalle, Ramesh, Qureshi, & Sankar, (2013) the contribution for each variable is determined by randomly permuting the values of that variable among the training points (both presence and background) and measuring the resulting decrease in training AUC. A large decrease in the AUC score indicates that the model heavily relied on that particular variable.

Jackknife tests produce an alternate estimate of variable importance in three different graphs. The first graph shows variable importance when each variable is excluded in turn, after which a model is created with the reaming variables, then again using each variable in isolation. The second and third graphs are of the test data and the AUC scores.

MaxEnt model AUC scores range from 0.89 – 0.912, all displaying relatively high predictability in reference to the 0.50 threshold. Models 1 and 4 (NEM1 & NEM4) produced the same AUC of 0.912 indicating the two models have similar predictive power. In both models annual temperature maximum had the highest variable importance of 83.8% and 89.1% respectively. The response curves for all models show how each variable affects the MaxEnt prediction. The red line represents the mean response of the 15 replicates; the blue is +/- one standard deviation. We can see that predictive suitability increases across all models with annual precipitation to generally 1250 mm then

decreases as precipitation continues to increase. Predictive suitability increase with annual temperature minimum as temperature increase to roughly 3.5 degrees Celsius, then the error margin becomes progressively more spread simulating a shotgun blast pattern. Similar to annual temperature minimum, the predictive suitability increases as annual temperature maximum increase to 16 degrees Celsius, after which the margin of error increases to resemble a similar shotgun pattern.

Looking at the first jackknife graph of model 1 (NEM1), we can see that annual precipitation is the least informative variable, while annual temperature maximum is the most informative variable. Results are similar in the second graph of the test data. Interestingly, looking at the jackknife of the AUC scores, we can see that if precipitation were to be left out, the AUC actually increases to slightly higher than what the full model produces.

Model 4 (NEM4) jackknife graphs show that annual temperature maximum is also the most informative variable. However, the AUC score is slightly higher when both variables are included in the model, vs. model 5 (NEM5) where only annual temperature maximum is included.

With the MaxEnt algorithm, we can say that annual temperature maximum is the most informative variable, followed by annual temperature minimum, while annual precipitation is the least informative variable. Based off the AUC scores, model 1 (NEM1) and model 4 (NEM4) have the same predictive power.

Figure 11 : MaxEnt Model 1 (NEM1) area under the curve of the reciever operating characteristic

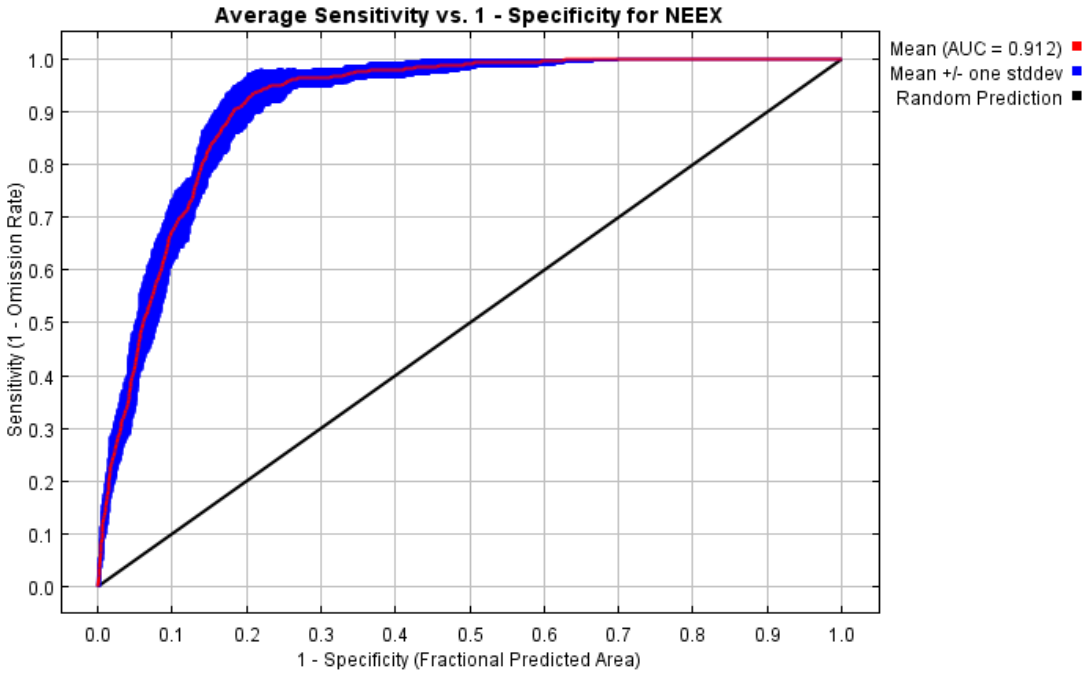
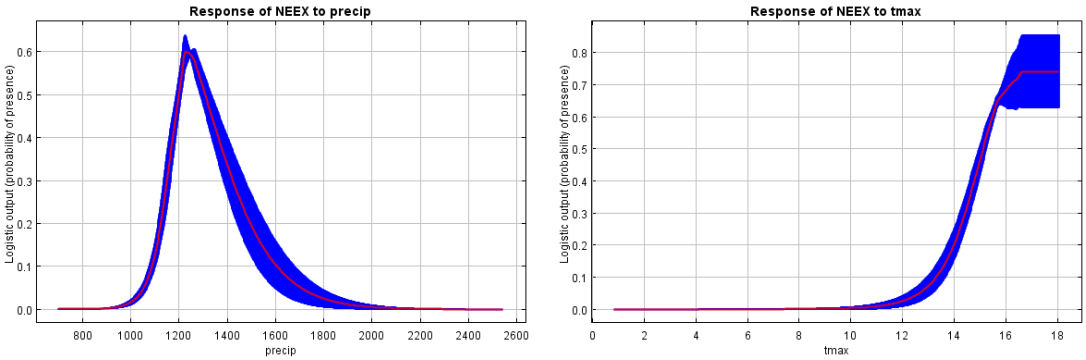


Table 5: Variable contribution table.

Variable	Percent contribution	Permutation importance
Temp Max	83.7	83.8
Temp Min	15.6	13.7
Precipitation	0.7	2.5

Figure 12 : MaxEnt Model 1 (NEM1) response curves of annual precipitation, annual temperature maximum, and annual temperature minimum.



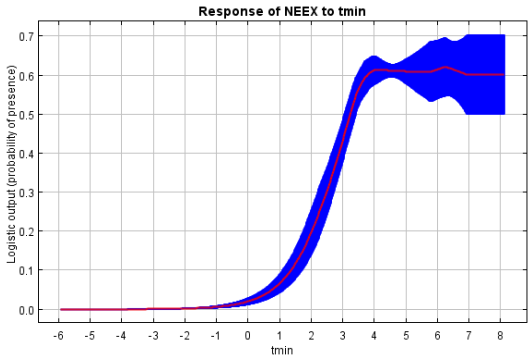


Figure 13 : MaxEnt Model 1 (NEM1) jackknife tests of model training data, test data, and AUC.

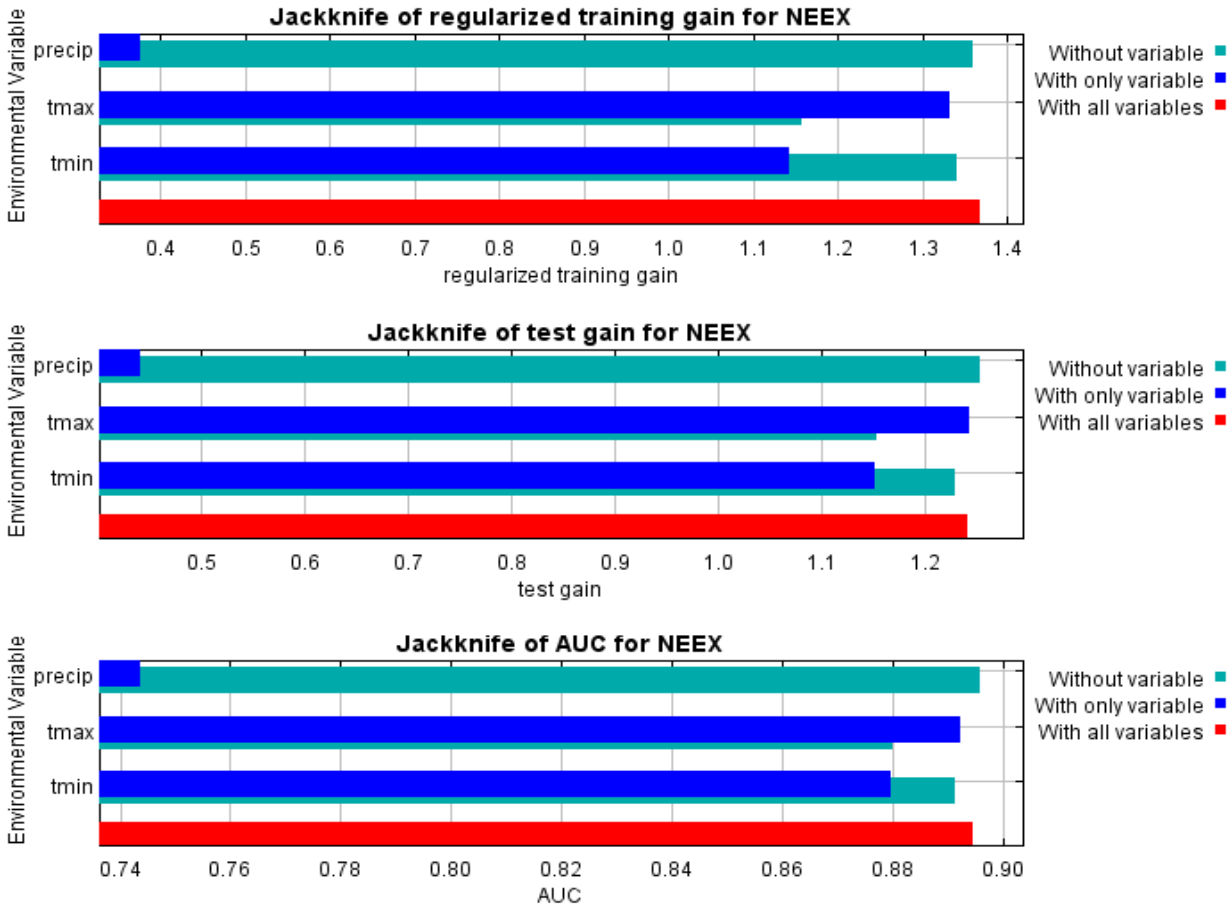


Figure 14: MaxEnt Model 1 (NEM1) raster habitat suitability map.

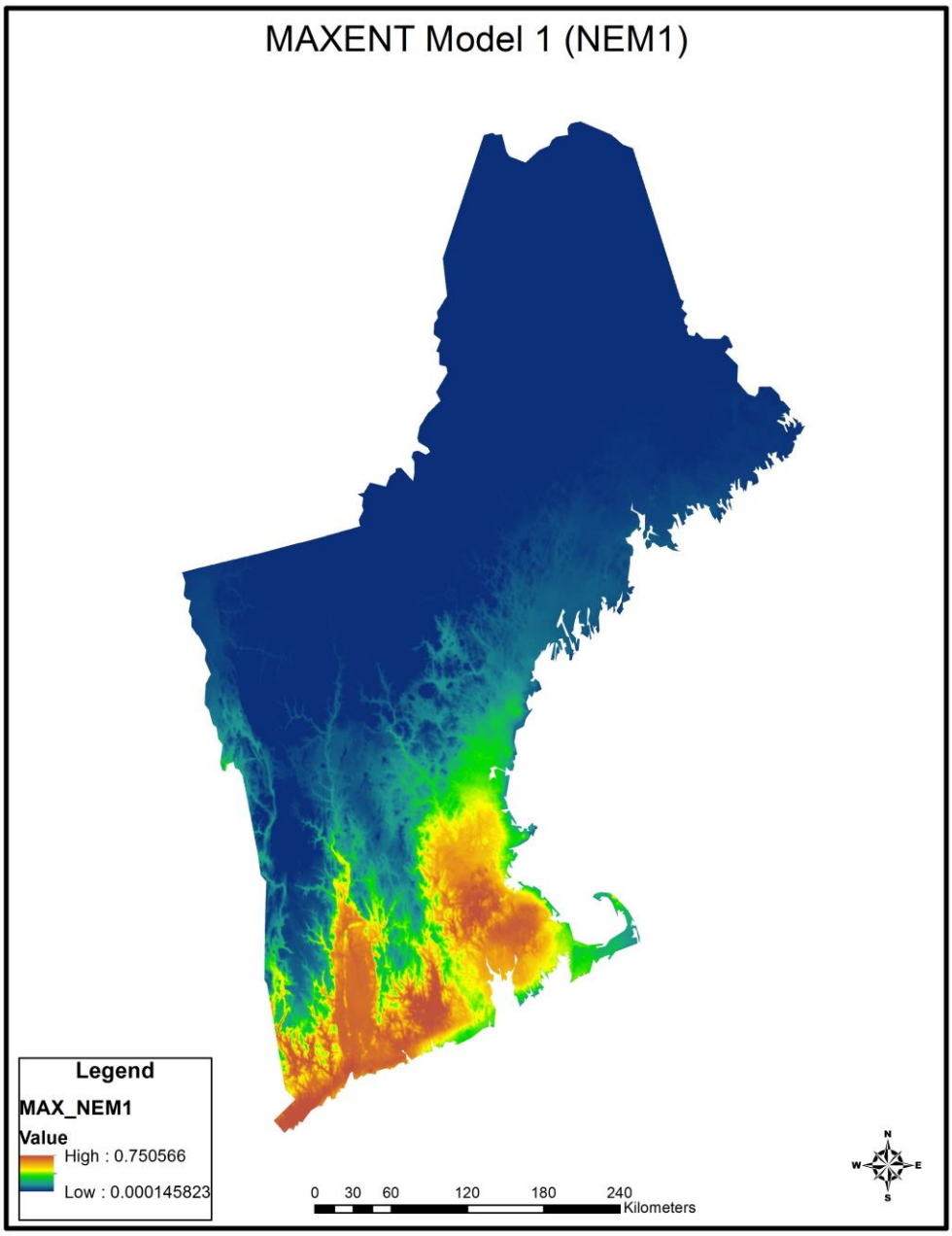


Figure 15: MaxEnt Model 4 (NEM4) area under the curve of the receiver operating characteristic.

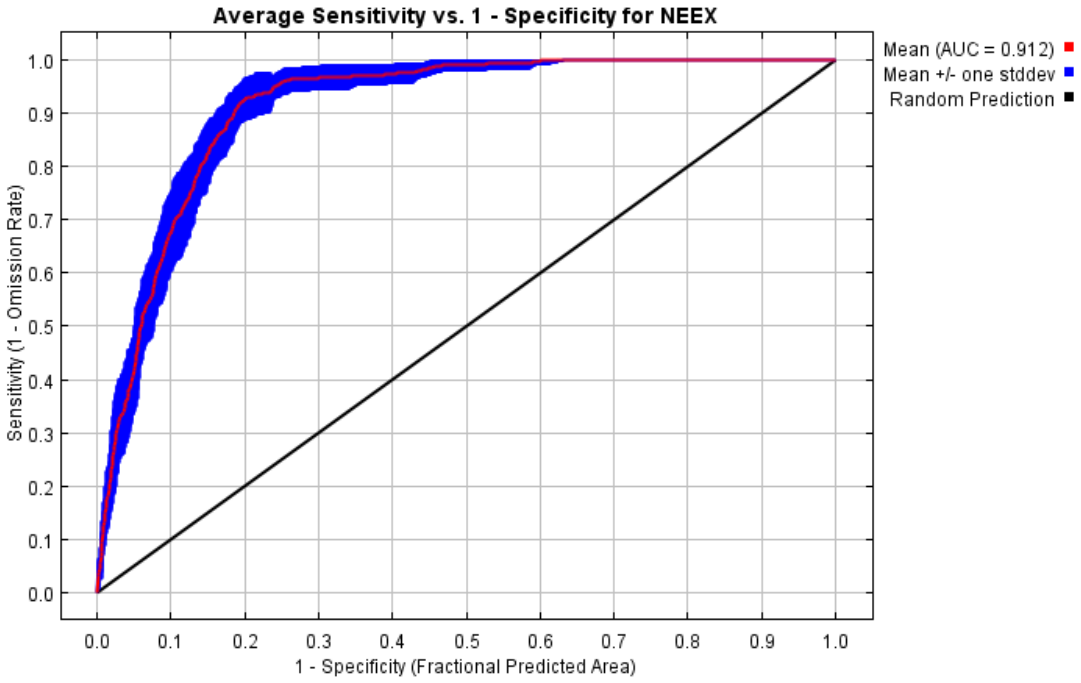


Table 6: Variable contribution table.

Variable	Percent contribution	Permutation importance
Temp Max	83	89.1
Temp Min	17	10.9

Figure 16: MaxEnt Model 4 (NEM4) response curve of annual temperature maximum and annual temperature minimum.

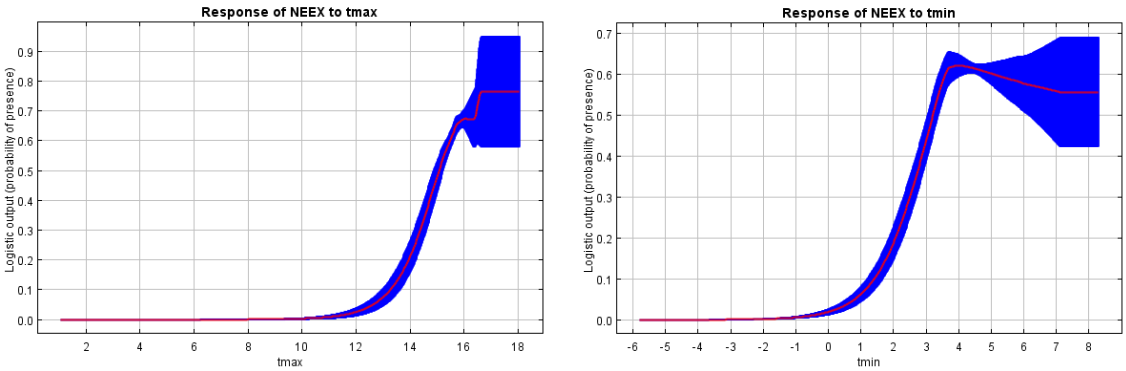


Figure 17: MaxEnt Model 4 (NEM4) jackknife test of model training data, test data, and AUC.

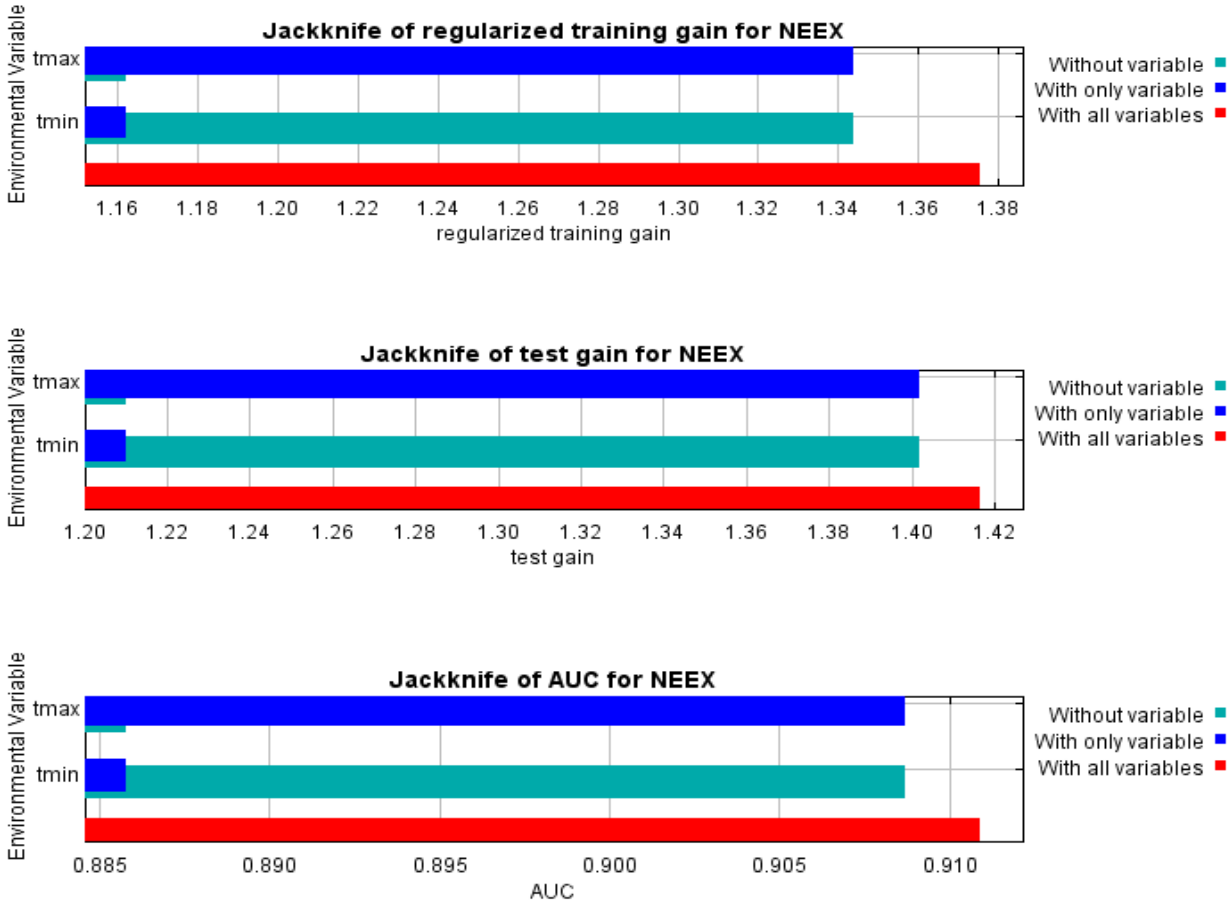
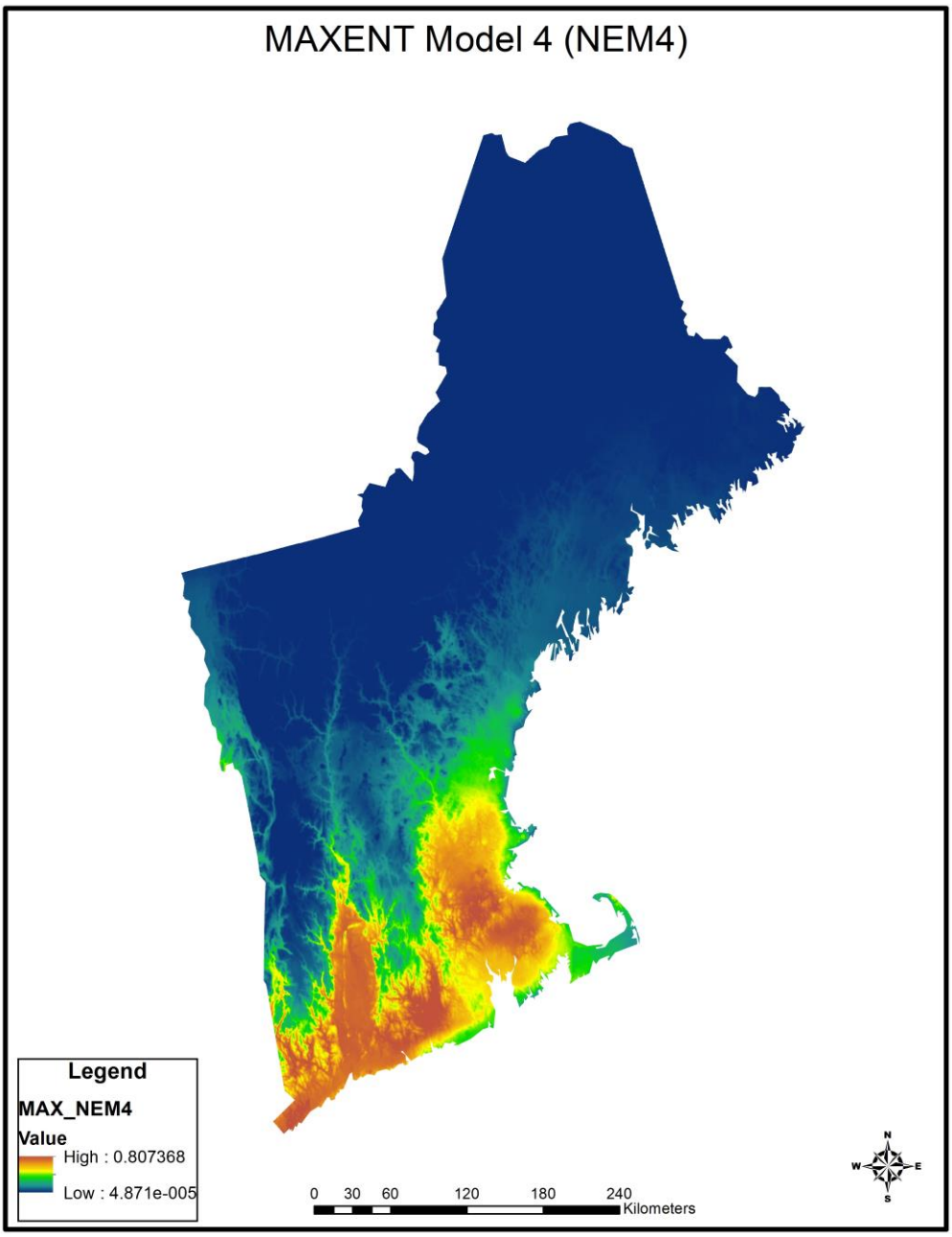


Figure 18: MaxEnt Model 4 (NEM4) raster habitat suitability map.

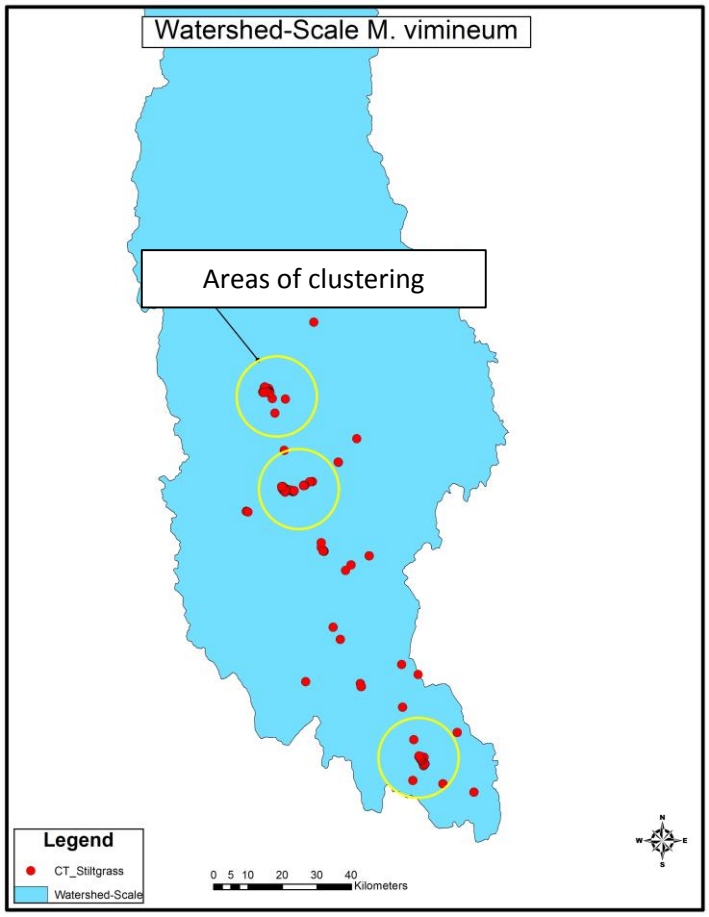


Watershed-Scale

Sample Point Evaluation

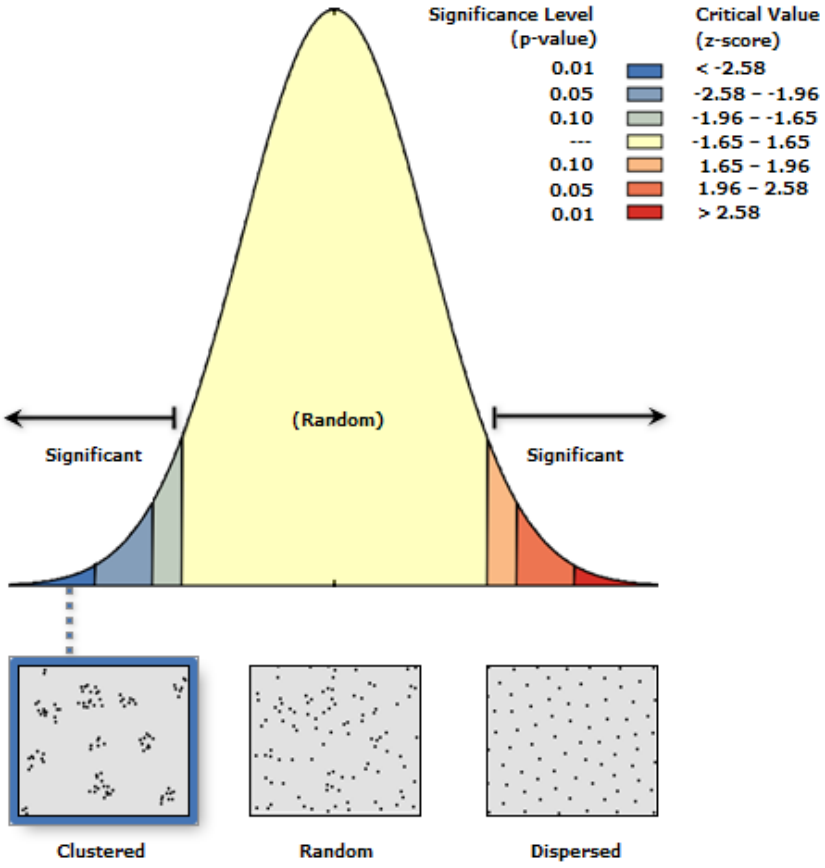
The same 1078 presence points used for sample point evaluation within the regional-scale were used for the watershed-scale. The points were then clipped to the Connecticut River watershed in similar fashion. Visual inspection of the 294 remaining points also revealed the possibility of sample clustering due to the sampling intensity around easily accessible areas (Figure 19).

Figure 19: Sample point evaluation revealing possible sample clustering.



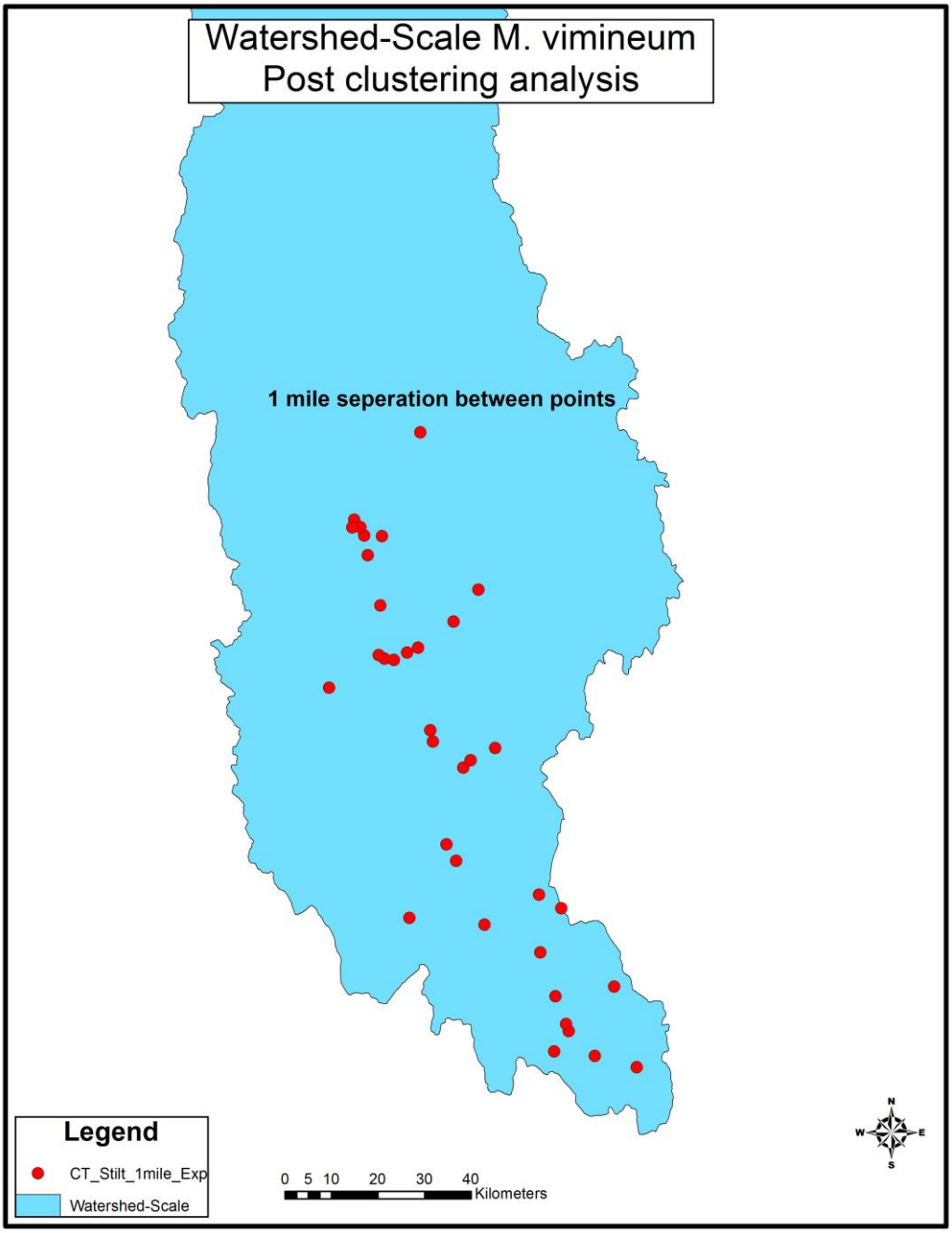
To reduce the adherent clustering, sample points were evaluated in the "Average Nearest Neighbor" tool in ArcGIS 10.2. Results from the nearest neighbor tool are as seen in Figure 20, the points were significantly clustered (P-value = 0.0000001).

Figure 20: Average nearest neighbor test confirming sample clustering.



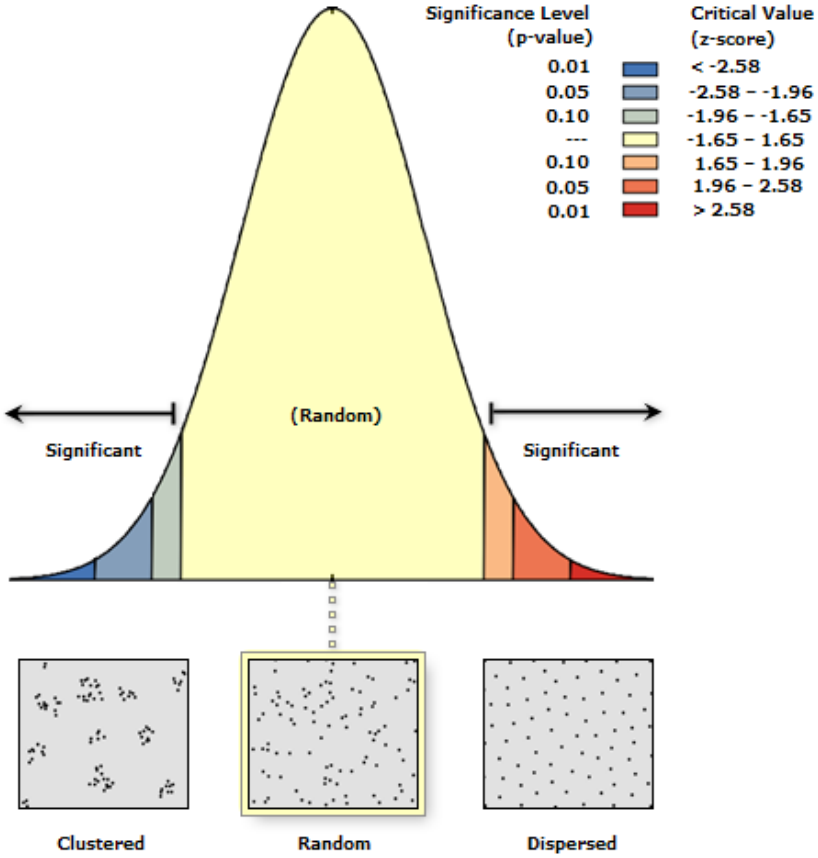
In an attempt avoid clustering; a point resample with a threshold of one mile was conducted in ArcGIS 10.2 (Figure 21).

Figure 21: Post-clustering analysis (1 mile separation between points).



35 sample points remained after the resample with no indication of clustering. P-value = 0.671164 (Figure 22).

Figure 22: Average nearest neighbor test revealing no evidence of clustering.



Variable Selection

Landscape and topographic variables were evaluated in the statistical program R using the biostats package designed by Dr. Kevin McGarigal of the University of Massachusetts, Amherst (McGarigal, 2013). A scatter plot matrix (SPLOM) with the Pearson's correlation coefficient (Figure 23) displays the non-clustered points and the variable's direct, indirect, or no evidence of correlation with one another. A cut-off value of 0.30 (positive or negative) was established to determine correlation. As a result, distance to hard features and topographic wetness was removed from further analysis as having the highest correlated values among all predictors. Solar radiance was highly correlated with aspect, therefore it was theorized that the amount of solar radiance a

surface receives is correlated with its compass orientation on the landscape i.e., southern-facing slopes receive higher amounts of solar radiance. Thus solar radiance was removed from further analysis. Five models were constructed a priori under the basic understanding of plant biology.

Model 1 (CTM1): Aspect, elevation, distance to water features, available water supply, and soil pH.

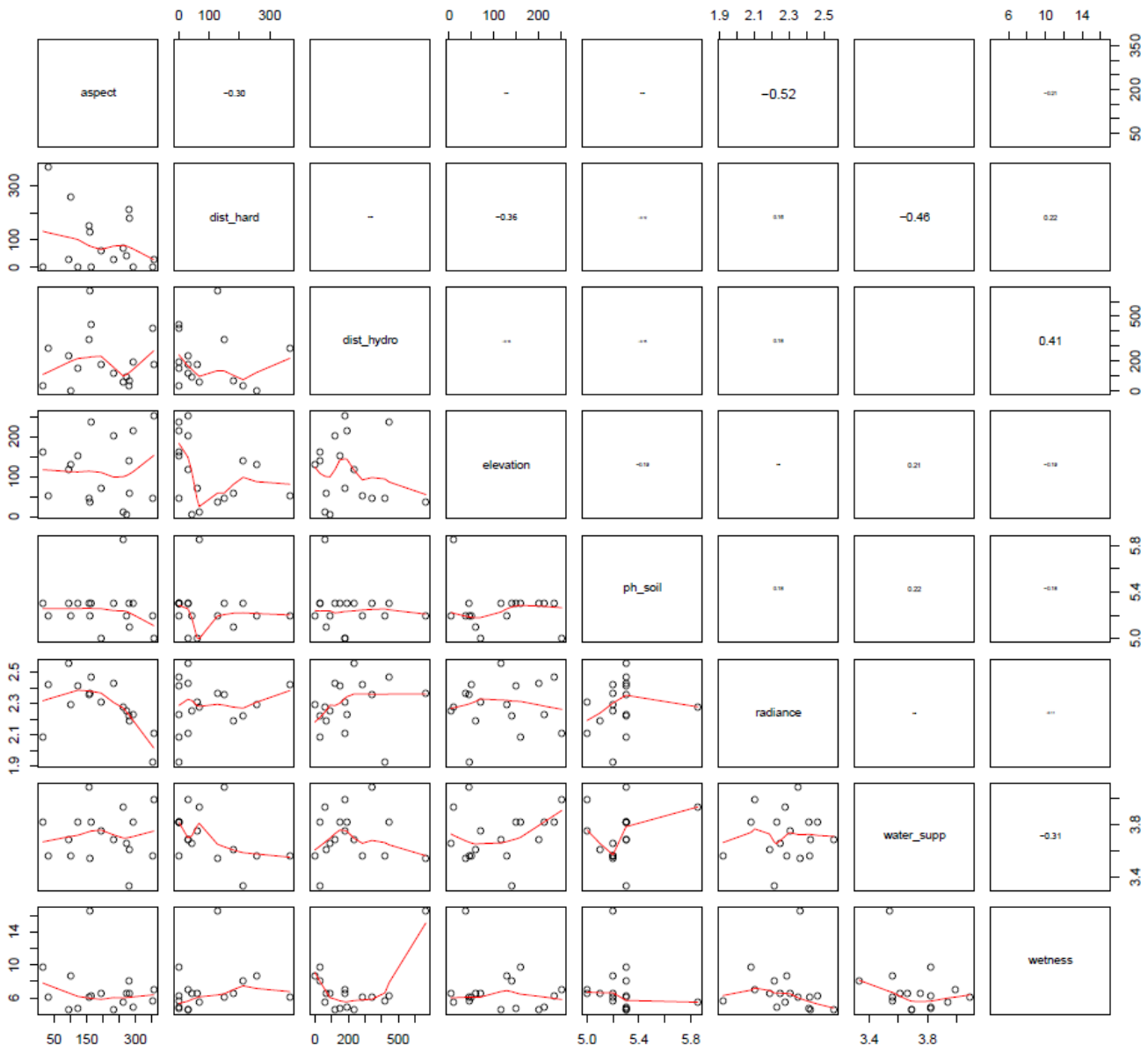
Model 2 (CTM2): Aspect, elevation, distance to water features, and soil pH.

Model 3 (CTM3): Elevation, available water supply, and soil pH.

Model 4 (CTM4): Elevation, and distance to water features.

Model 5 (CTM5): Elevation

Figure 23: Scatter plot matrix of predictor variables for the watershed-scale and associated Pearson's coefficients.



Models

Generalized Linear Model

The GLMs were fitted in R software version 2.15.1. 10000 pseudo-absence points were randomly generated in ArcMap 10.2 using the “create random points” tool with a

minimum distance of one meter between each of the points and restricted any random points to fall on actual presence points.

All 35 presences and pseudo-absence points were "merged" together and the predictor cell values where points existed were extracted using the "extract multiple values to points" tool in ArcGIS 10.2. The attribute table was then exported as a .CSV file compatible with R.

Ecological Niche Factor Analysis

The same groups of variables and the same 35 presence points used in the GLM were applied to the five ENFA models. OpenModeller software was used to run ENFA and the default was accepted for all model parameters.

Maximum Entropy Algorithm

Like the GLM and ENFA models, the 5 groupings of variables and the 35 presence points were applied to the MaxEnt models. Each of the five model's parameters was adjusted to allow for validation, replication, and optimization. 30 percent of the presence localities were set aside in a random seed method to be used for model validation. The number of replicates was increased from 1 to 15 to allow for more averaging across model runs. Replicated run type was set to the bootstrap method of sample replacement. The training iterations were increased from 500 to 5000 for more optimization. The default was accepted for all other parameters.

Results

Results from the GLM suggest that model 4 (CTM4) distance to water features and elevation were the most significant predictor (Table 7). The model also had the lowest AIC score of 389.48 (Table 8). Since the deltaAIC scores were relatively low among most of the models, a weighted model averaging was conducted (Table 8). We

can see that model 4 (CTM4) contributed 52% of the explained variance. However, when distance to water features is removed from model 4 like in model 5 (CTM5), contribution lowered by 27%. Elevation remained the only significant variable among all models. Distance to water features displayed a trend towards significance, however remained only significant at the 0.10 p-value level.

To spatial display the results from the GLM models, each model's formula were scripted into the "Raster calculator" tool in ArcGIS 10.2. The resulting rasters were scaled from 0-1 displaying actual probability of occurrence of *M. vimineum*.

Table 7: GLM model 4 (CTM4) outputs (formula = Abundance ~ elevation + dist_water, data = CTPA, family = binomial)

	Estimate	Std. Error	Z Value	Pr(>|z|)
Intercept	−3.019	0.290	-10.401	$< 2e-16$ ***
Elevation	−0.011	0.002	-5.849	$4.96e-09$ ***
Dist_water	-0.002	0.001	-1.712	0.087
AIC: 389.484				

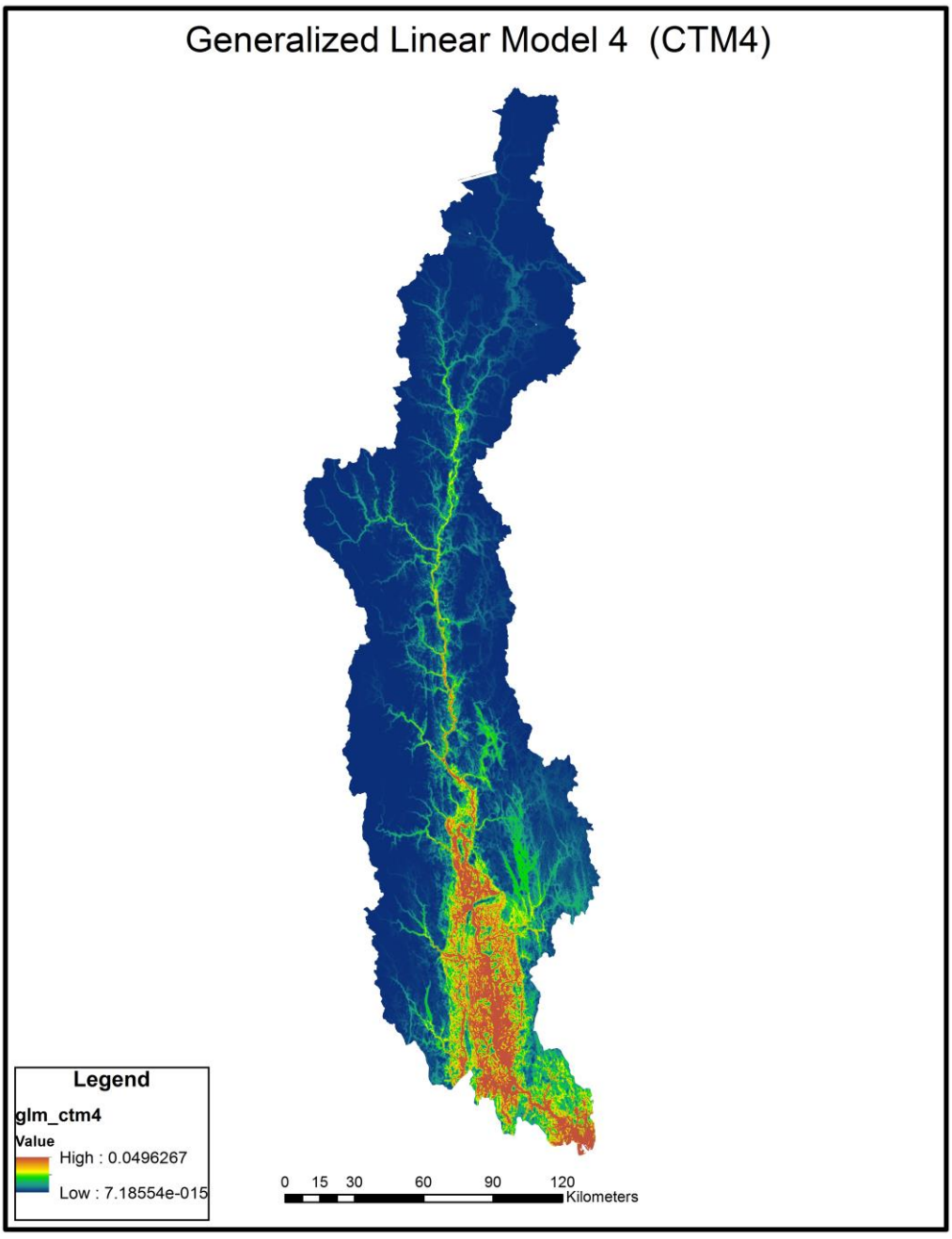
Table 8: Akaike information criteria (AIC) table.

Model	DF	AIC
CTM4	3	389.484
CTM5	2	390.969
CTM2	5	392.754
CTM1	6	393.629
CTM3	4	393.667

Table 9: Model averaging components table.

Variable	DF	LogLik	AIC	Delta	Weight
23	3	−191.74	389.49	0.00	0.52
3	2	−193.48	390.97	1.48	0.25
1234	5	−191.38	392.76	3.27	0.10
12345	6	−190.81	393.64	4.15	0.07
345	4	−192.83	393.67	4.19	0.06
TERM CODES:	**Aspect**	**Distance to water**	**Elevation**	**Soil pH**	**Soil AWS**
	1	2	3	4	5

Figure 24: GLM Model 4 (CTM4) raster probability of occurrence map.



ENFA models 1 (CTM1), 2 (CTM2), 3 (CTM3), and 4 (CTM4) have all appear to fail and are not evaluated as significant models. Although the models produced marginality and specialization factors, the AUC score were 0.50, or no better than

random. This could be a result of the failure of the broken stick discard method due to sampling not covering a large enough range of component values, but a definite cause for model failure was not achieved.

Model 5 (CTM5) incorporated only one variable, elevation, which of explained 1.00 or 100% of the variance and produced an AUC score of 0.60. With a marginality factor 0.65 and specialization factor of 2.44, model 5 (CTM5) suggest that *M. vimineu's* distribution is fairly general within the study area, but highly specialized among the component, in this case elevation. The model output raster for CMT5 confirms that *M. viminuem* is distributed among the lower lying areas within the study area.

Figure 25: ENFA Model 5 (CTM5) Area under the curve of the reciever operating characteristic.

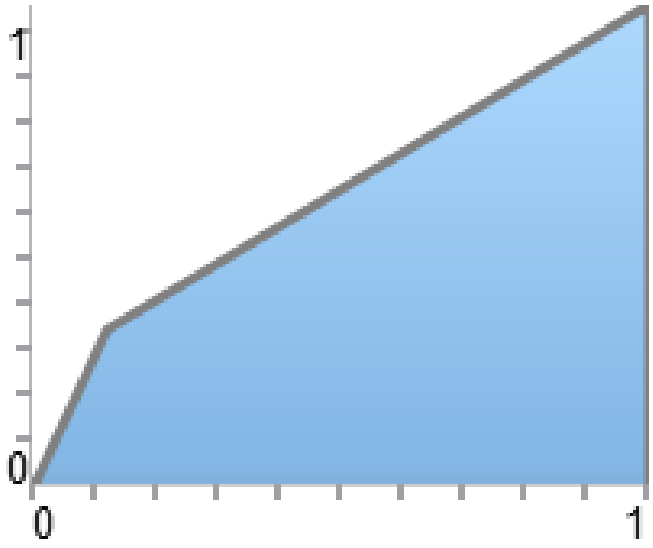
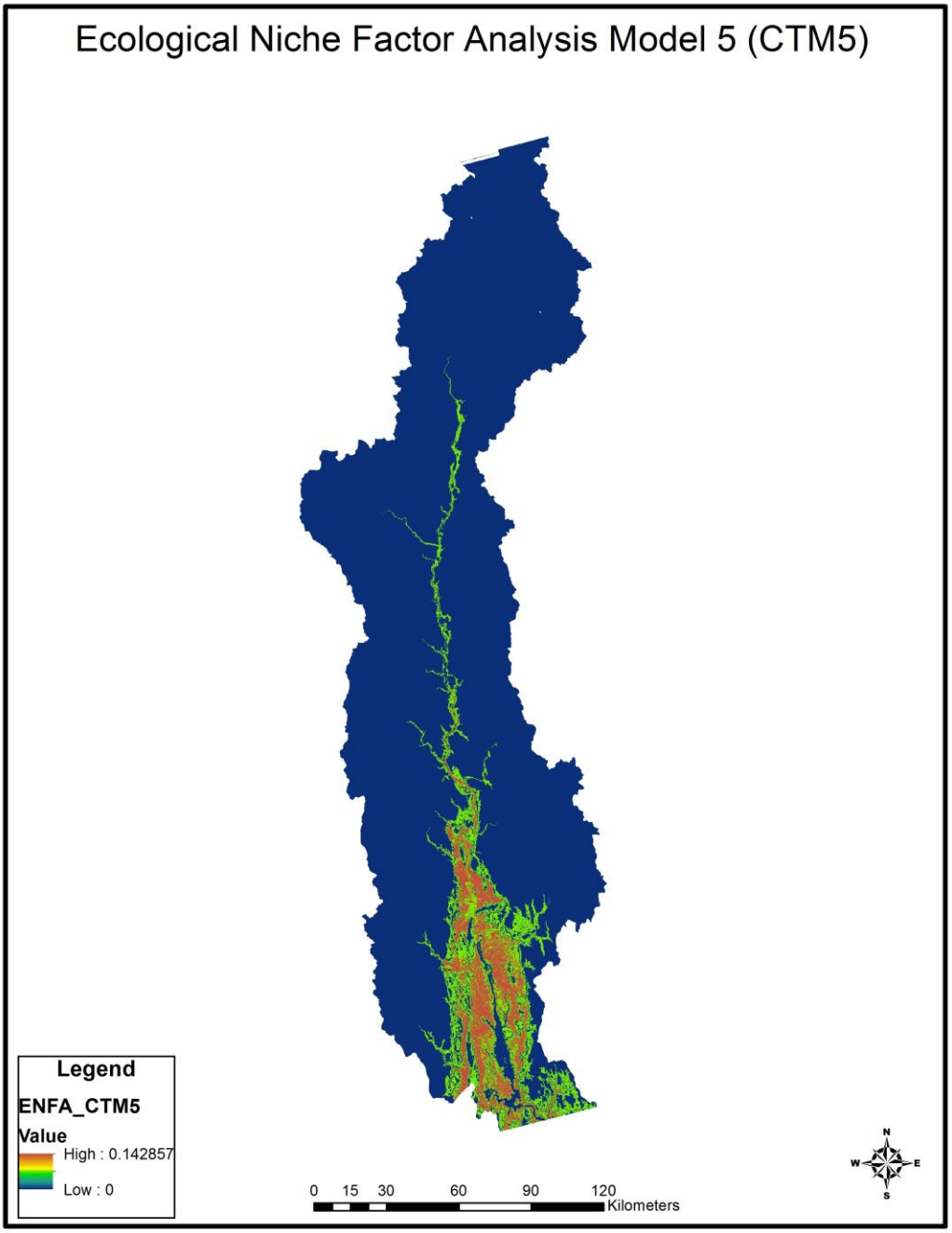


Figure 26: ENFA Model 5 (CTM5) raster habitat suitability map.



MaxEnt model AUC scores range from 0.88 – 0.935, all displaying relatively high predictability in reference to the 0.50 threshold. Model 1 (CTM1) produced an AUC of 0.935 indicating the highest predictive power, as appose to model 5 (CTM5) of ENFA.

The response curves for all models show how each variable affects the MaxEnt prediction. The red line represents the mean response of the 15 replicates; the blue is +/- one standard deviation. We can see that predictive suitability decreases across all models as elevation increases, and as the distance to water features increase. *M. vimimeum's* response to aspect appears to be very general, indicating that this species can tolerate a wide range of orientation, however prefers between 50 – 350 topographic orienting. The response to soil available water supply tends to have an hourglass shape that becomes increasingly spread as water is made more available. The probability of presence decreases after more than 3.5 cm of volumetric water is available within the soil. *M vimineum's* response to soil pH seems to be unlike other grasses which prefer pH levels in the 6.0 – 6.5 range. *M. vimineum* appears to prefer slightly poor soil pH levels (5.1 – 5.8) which is in-line with U.S. Forest Service findings (USDA, 2015).

Model 1 (CTM1) variable permutation importance test indicates that elevation is by far the most important variable (72.7) followed by distance to water features (8.9) and soil pH being the least important (4.9). Elevation remained the most important variable across all models, however distance to water features varied within models.

Looking at the first jackknife graph of model 1 (CTM1), we can see that elevation is the most informative variable, while aspect is the least informative variable, which is different than the permutation importance. Results are similar in the second jackknife graph of the test data. Interestingly, looking at the jackknife of the AUC scores, we can see that if soil available water supply were to be left out, the AUC would actually increase to slightly higher than what the full model produces. This indicates the existence of a set variables that possess greater predictive power.

With the MaxEnt algorithm, we can say that elevation is the most informative variable, followed by soil pH and distance to water features, while aspect and soil available water supply are the least informative variables.

Figure 27: MaxEnt Model 1 (CTM1) area under the curve of the receiver operating characteristic

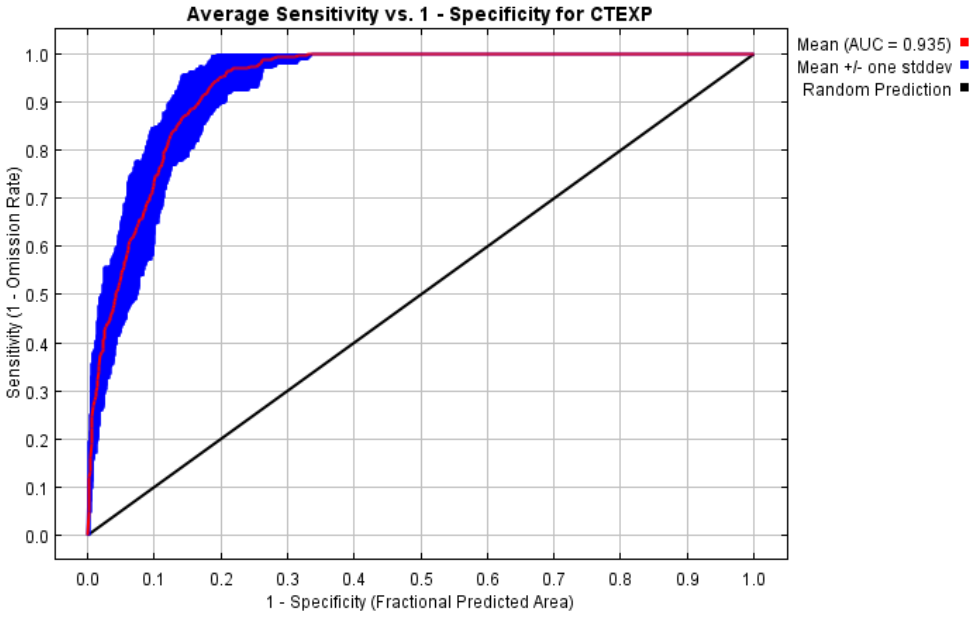


Table 10: Variable contribution table.

Variable	Percent contribution	Permutation importance
dem1	65.6	72.7
ph_soil	14.4	4.9
hydro_final1	9.6	8.9
Aspect	8.3	6.9
Aws	2	6.6

Figure 28: Response curves for elevation, soil pH, distance to water features, aspect, and soil available water supply

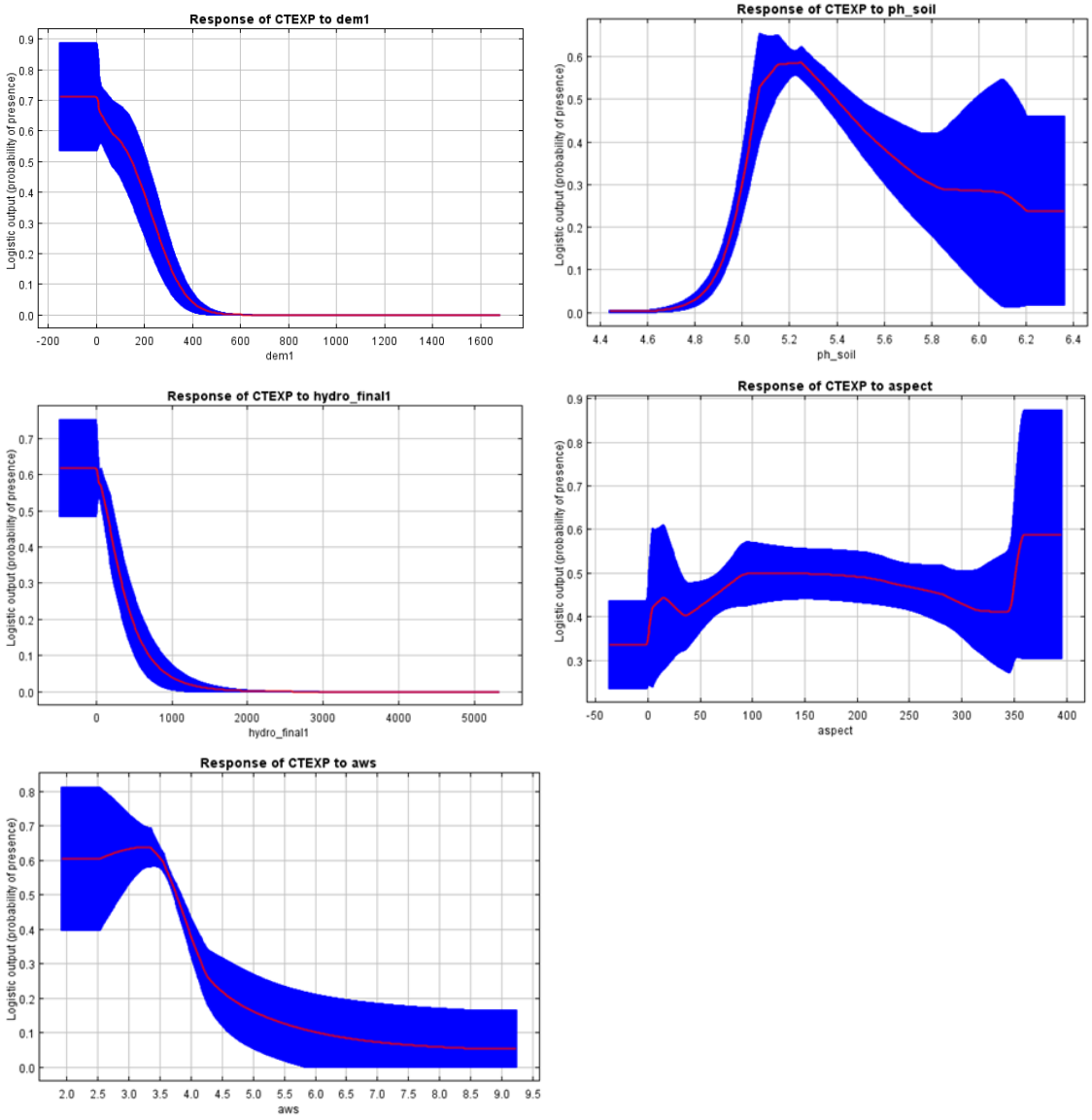


Figure 29: Jackknife graphs of training and test data, and AUC.

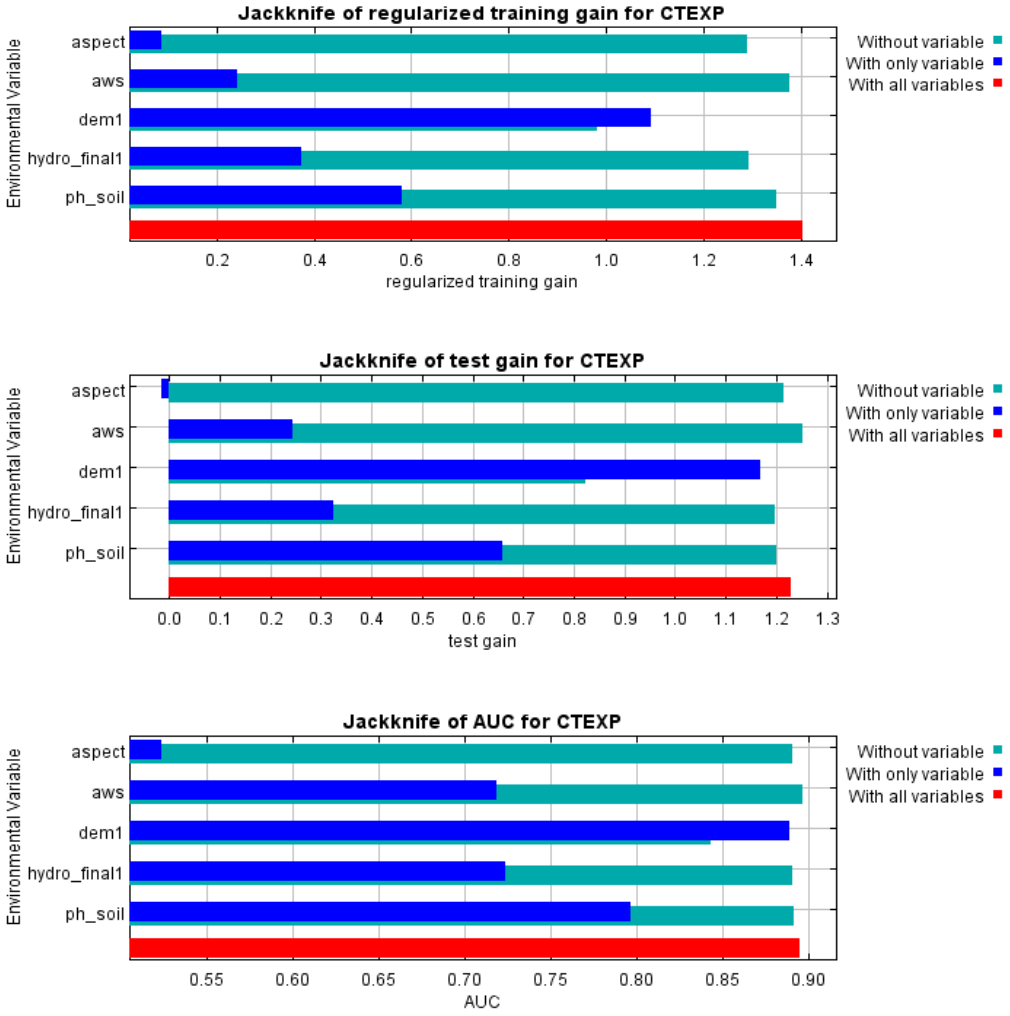
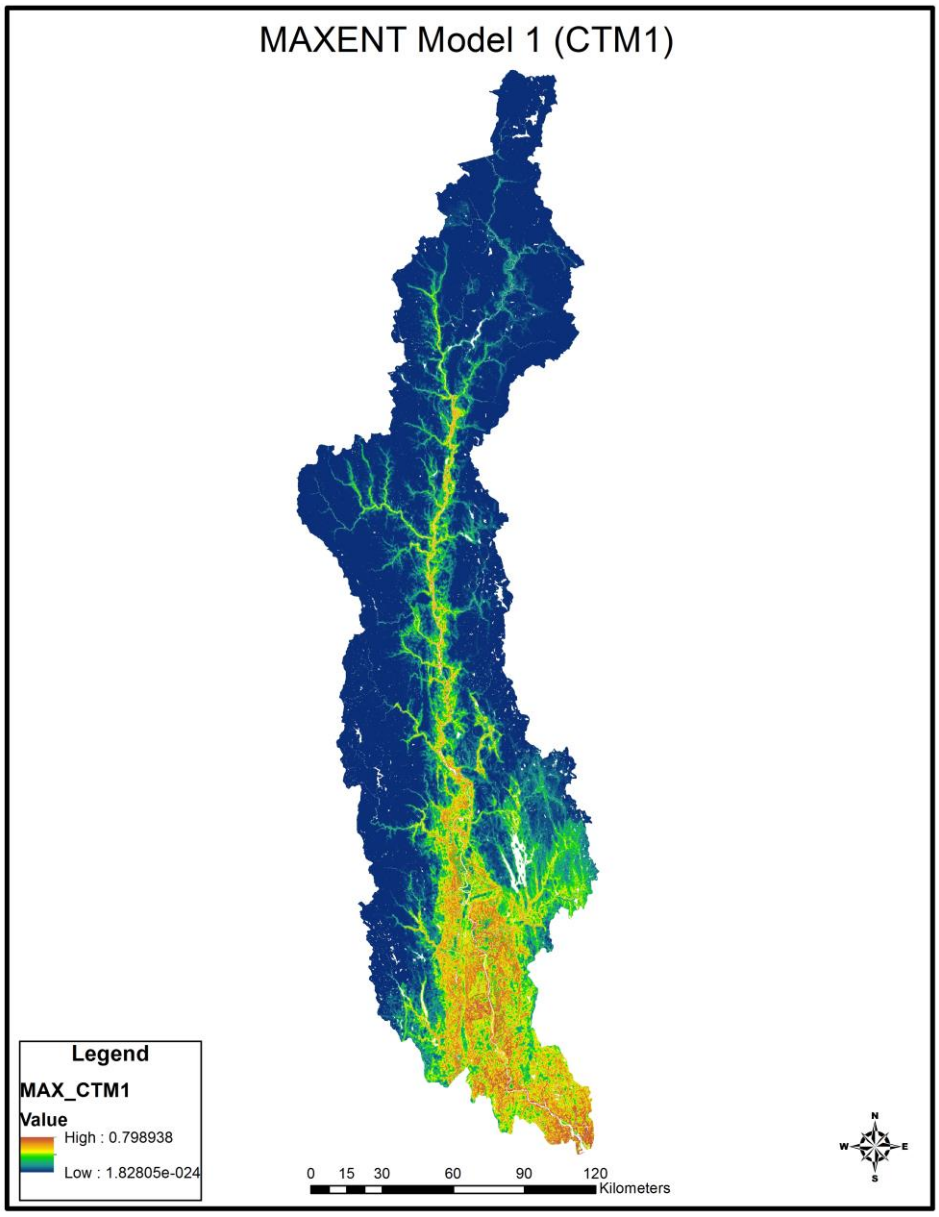


Figure 30: Raster map of MaxEnt Model 1 (CTM1)



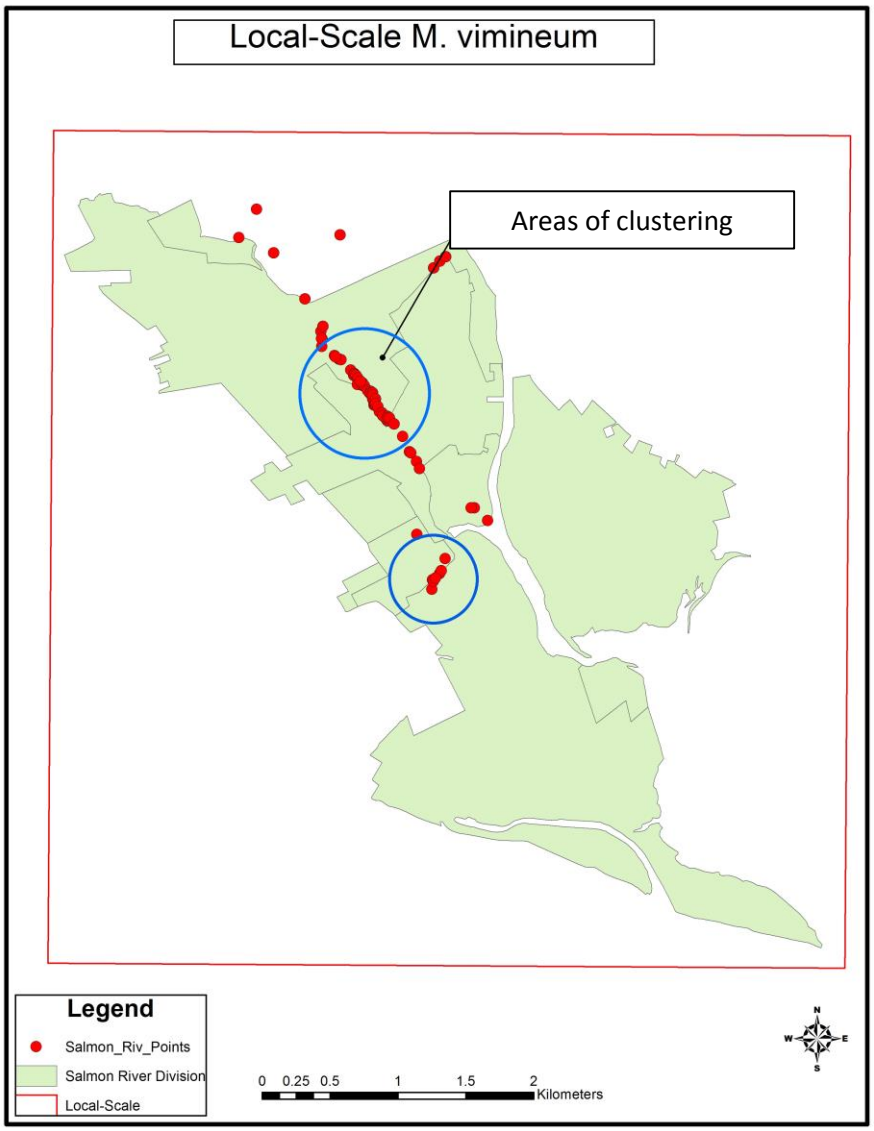
Local-Scale

Sample Point Evaluation

The same 1078 presence points used for sample point evaluation within the regional-scale and Watershed-scale were used for the local-scale. The points were then

clipped to the local-scale in a similar fashion. Visual inspection of the 106 remaining points also revealed the possibility of sample clustering due to the sampling intensity around easily accessible areas (Figure 31).

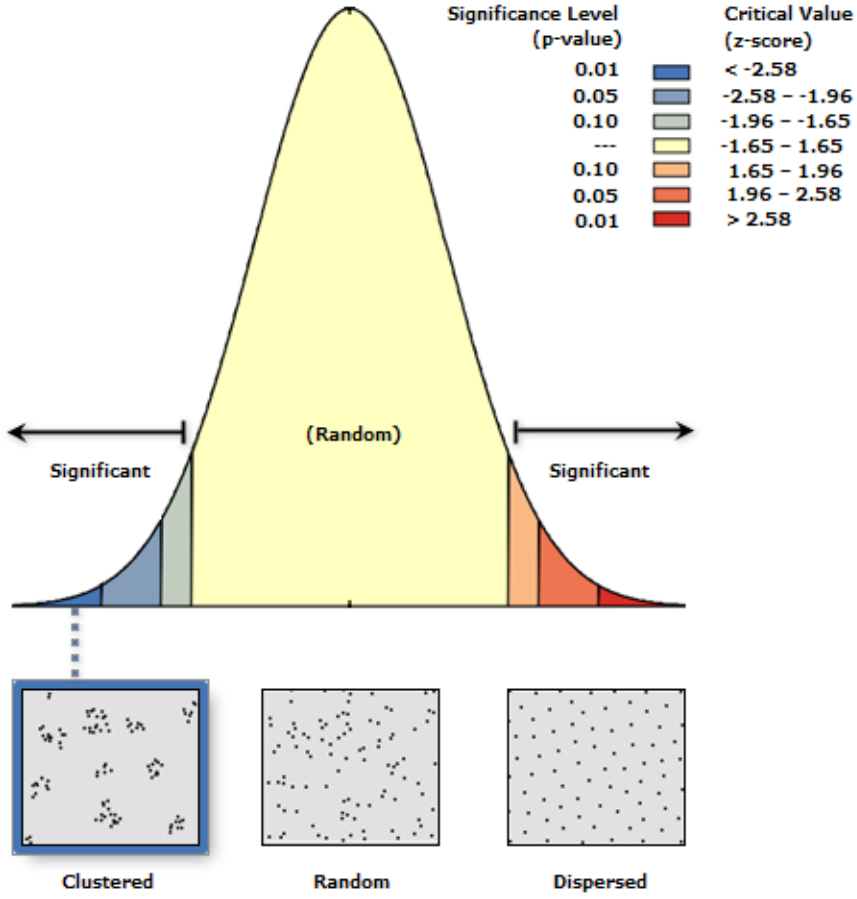
Figure 31: Local-scale areas of potential sample clustering.



To reduce the adherent clustering, sample points were evaluated in the "Average nearest neighbor" tool in ArcGIS 10.2. Results from the nearest neighbor tool are as seen

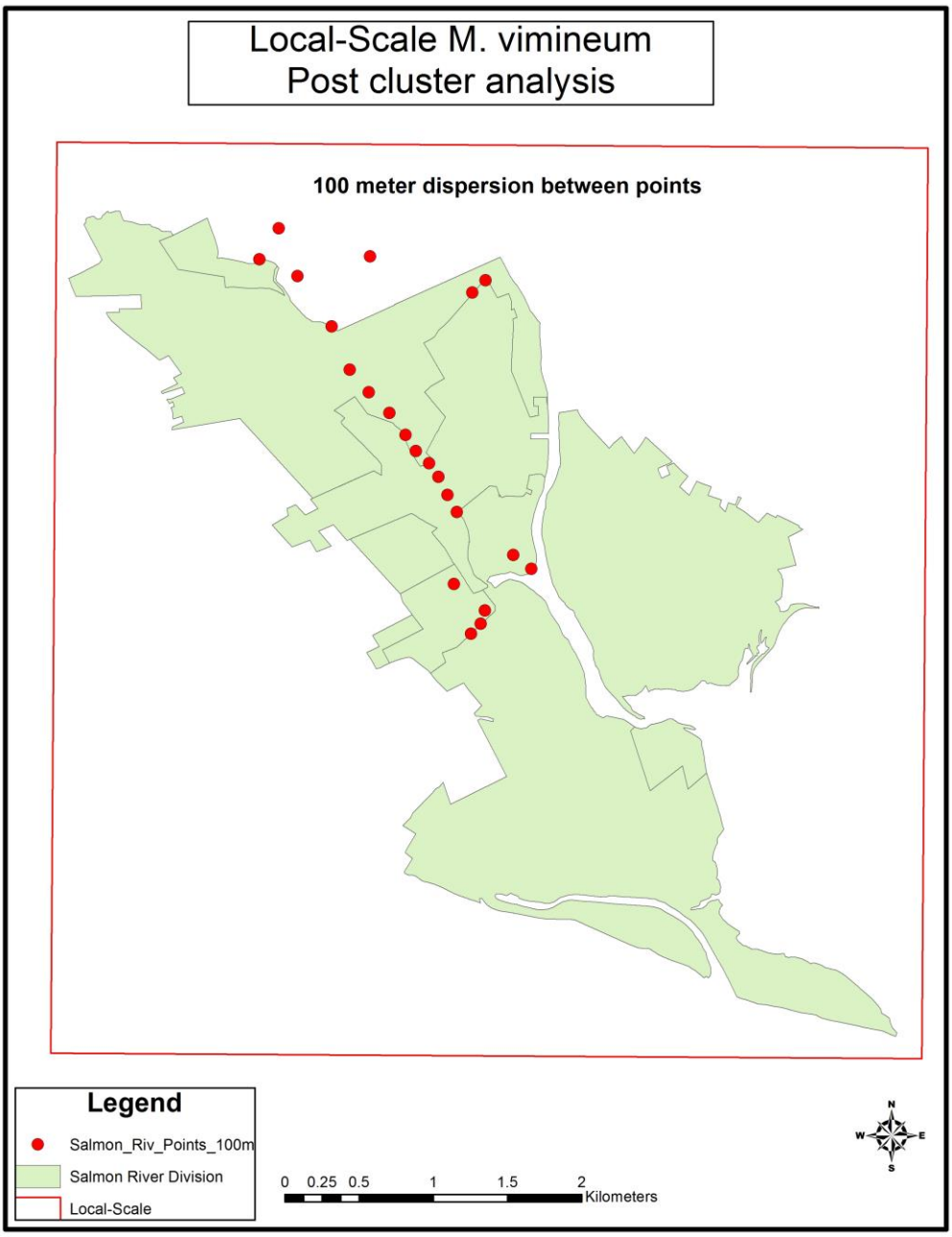
in Figure 32, the points were significantly clustered with a P-value = 0.00000001 and Z-score -16.1231

Figure 32: Average nearest neighbor analysis of local-scale sample locations.



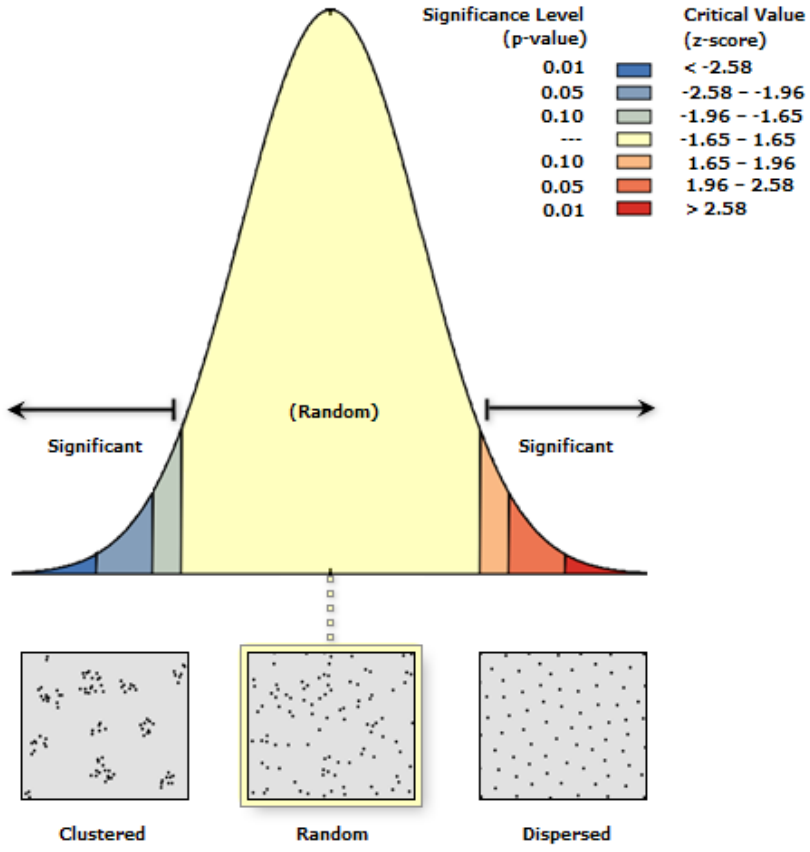
In an effort avoid sample clustering; a point resample with a threshold of 100 meters separation between points was conducted in ArcGIS 10.2 (Figure 33).

Figure 33: Map displaying 100 meter separation between sample points.



22 sample points remained after the resample and there was no evidence of clustering. As seen in Figure 34, the P-value = 0.2933 reveals that the point pattern is not significantly different than random.

Figure 34: Average nearest neighbor analysis on post-clustered sample points.



Variable Selection

Since the local-scale was analyzed at the one meter scale, only one predictor variable was available for analysis. NDVI was constructed from the 4-band NAIP imagery. Results from the NDVI produce values on a continuous scale from -1 to 1. No other one meter resolution freely available datasets were available for analysis.

Models

Generalized Linear Model

The GLM was fitted in R software version 2.15.1. Given the reduced size of the local-scale in comparison to other experimental scales, 1000 pseudo-absence points were randomly generated in ArcMap 10.2 using the "create random points" tool with a

minimum distance of one meter between each point and restricted any random points to fall on actual presence points.

All 22 presence and 1000 pseudo-absence points were "merged" together and the predictor cell values where points existed were extracted using the "extract multiple values to points" tool in ArcGIS 10.2. The attribute table was then exported as a .CSV file compatible with R.

Ecological Niche Factor Analysis

The predictor variable NDVI and the same 22 presence points used in the GLM were applied to the ENFA model. The default was accepted for all other model parameters.

Ecological niche factor analysis Model 1 (LM1) continually failed after several attempts. The ENFA algorithm could not calculate the square root for the matrix of the NDVI raster values where the raster values were negative. To correct the issue, one was added to all raster values. The results from the addition were then divided by two to achieve a positive value in all raster values.

Maximum Entropy Algorithm

Like the GLM and ENFA models, the NDVI variable and the 22 presence points were applied to the MaxEnt model. Model parameters were adjusted to allow for validation, replication, and optimization. 30 percent of the presence localities were set aside in a random seed method to be used for model validation. The number of replicates was increased from 1 to 15 to allow for more averaging across model runs. Replicated run type was set to the bootstrap method of sample replacement. The training iterations were increased from 500 to 5000 for more optimization. The default was accepted for all other parameters. Model 1 (LM1): NDVI.

Like the Ecological niche factor analysis, the MaxEnt Model 1 (LM1) continually failed after several attempts. The MaxEnt algorithm could not calculate the square root for the matrix of the NDVI raster values where the raster values were negative. To correct the issue, one was added to all raster values. The results from the addition were then divided by two to achieve a positive value in all raster values.

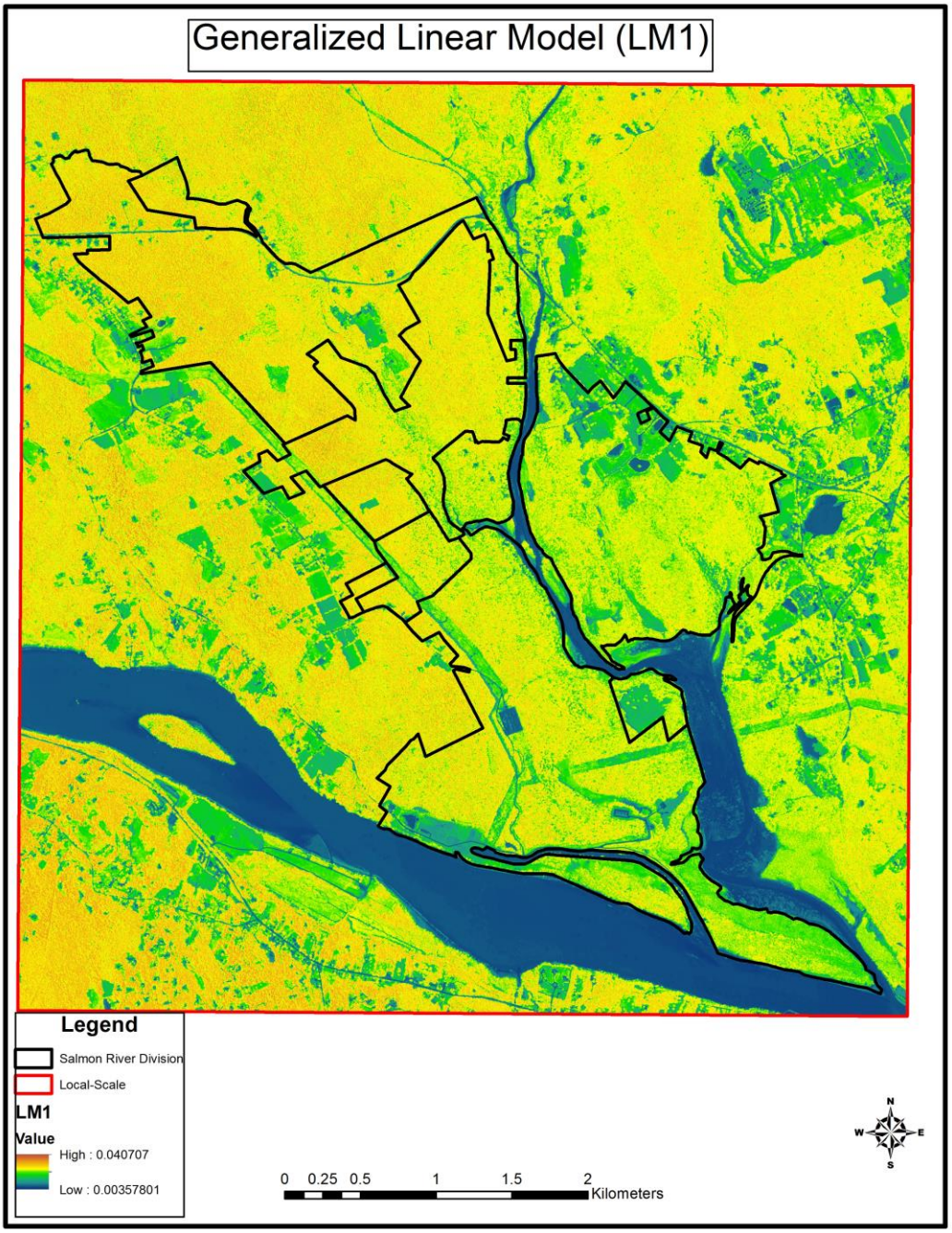
Results

Since there was only one model, there was no need to create a model selection or averaging tables. Results from the GLM indicate that NDVI is not a significant predictor of *M. vimineum* within the Local-scale. A nonsignificant result could be due to the fact that the sample size had been greatly reduced to avoid clustering and or sampling, although was random based off the clustering analysis, might not cover enough diverse NDVI values.

Table 11: GLM model 1 (LM1) outputs. Formula = Abundance ~ NDVI, family = binomial, data = Local.pa).

	Estimate	Std. Error	Z Value	Pr(>|z|)
Intercept	-4.147	0.345	-12.030	$< 2e-16$ ***
NDVI	1.650	1.117	1.478	0.139

Figure 35: Map displaying GLM model 1 (LM1) results.



The results from the ENFA suggest that the model actually did worse that what would have been predicted by random chance. Model 1 (LM1) produced an AUC score of 0.46, indicating less predictive power than a random prediction. Interpreting the

marginality factor of 0.12 and specialization factor of 1.54 given NDVI, *M. vimineum* is greatly found within the study area, but fairly specialized among NDVI values.

Figure 36: ENFA model 1 (LM1) area under the curve of the receiver operating characteristic.

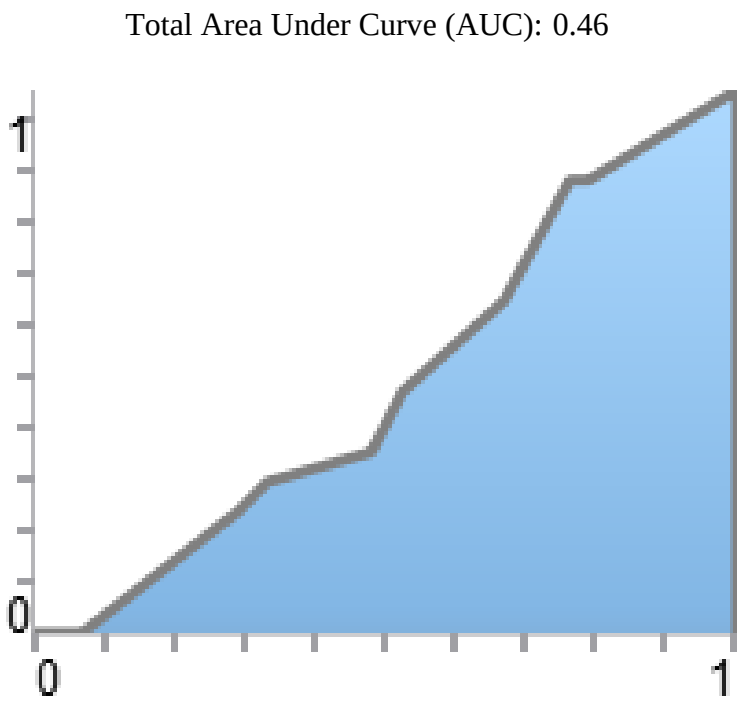
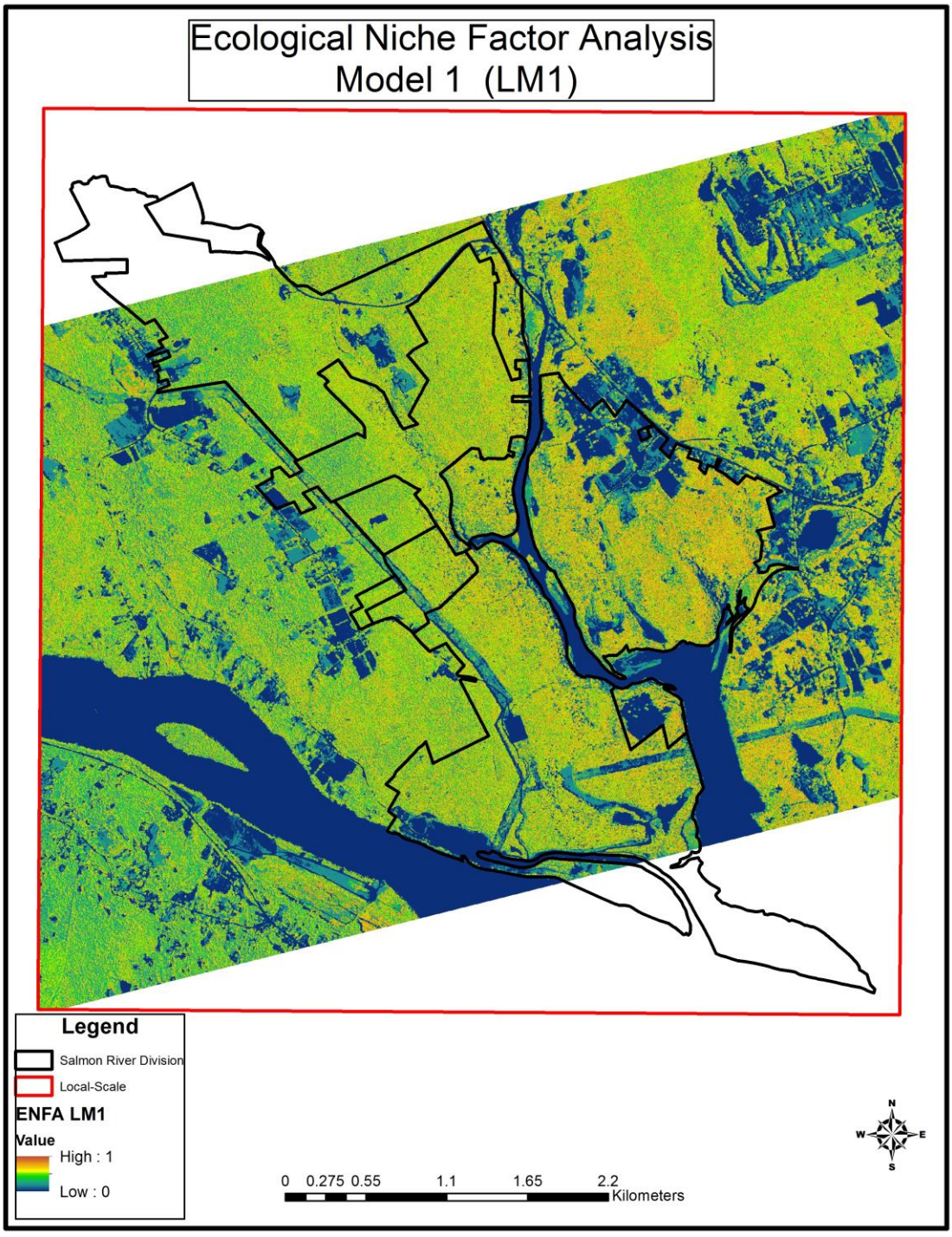


Figure 37: Map displaying the ENFA model 1 (LM1) results.



Like the ENFA algorithm, the MaxEnt model (LM1) produced a very low AUC of 0.56, which indicates no difference in predictive power than random. However, interpreting the response curve, we can see that *M. vimineum* covers a wide range of

NDVI values oppose to the relatively high specialization value in the ENFA which indicate a narrow niche. Although there is a wide range of NDVI values, the probability is only near 0.50 across those values.

Figure 38: MaxEnt model 1 (LM1) area under the curve of the receiver operating characteristic.

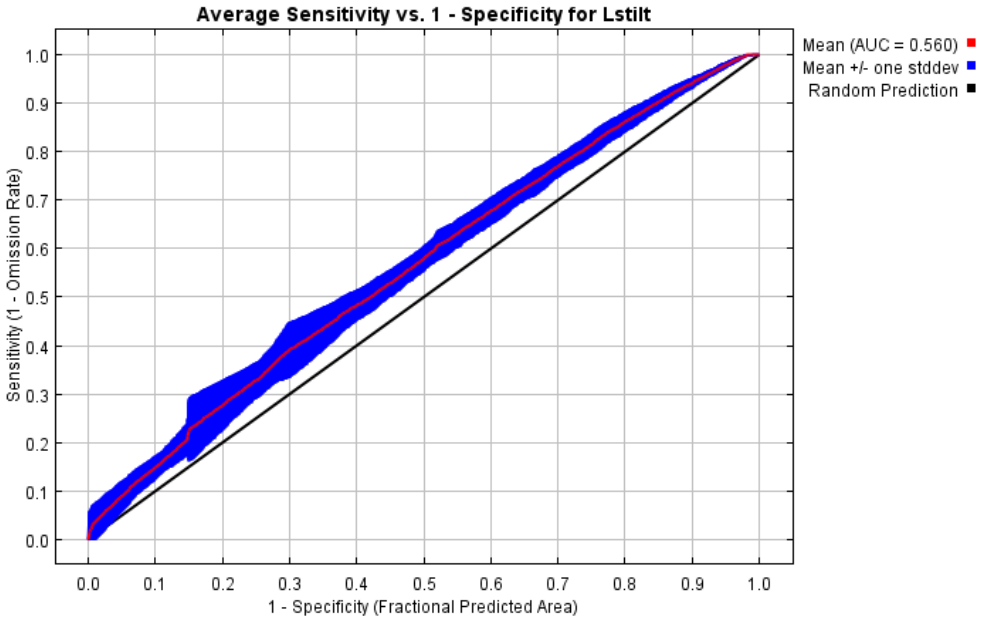


Figure 39: Response curve for NDVI.

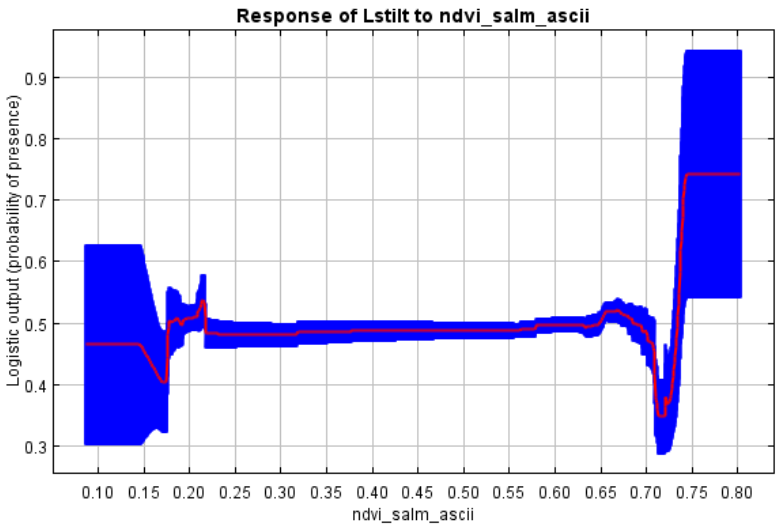
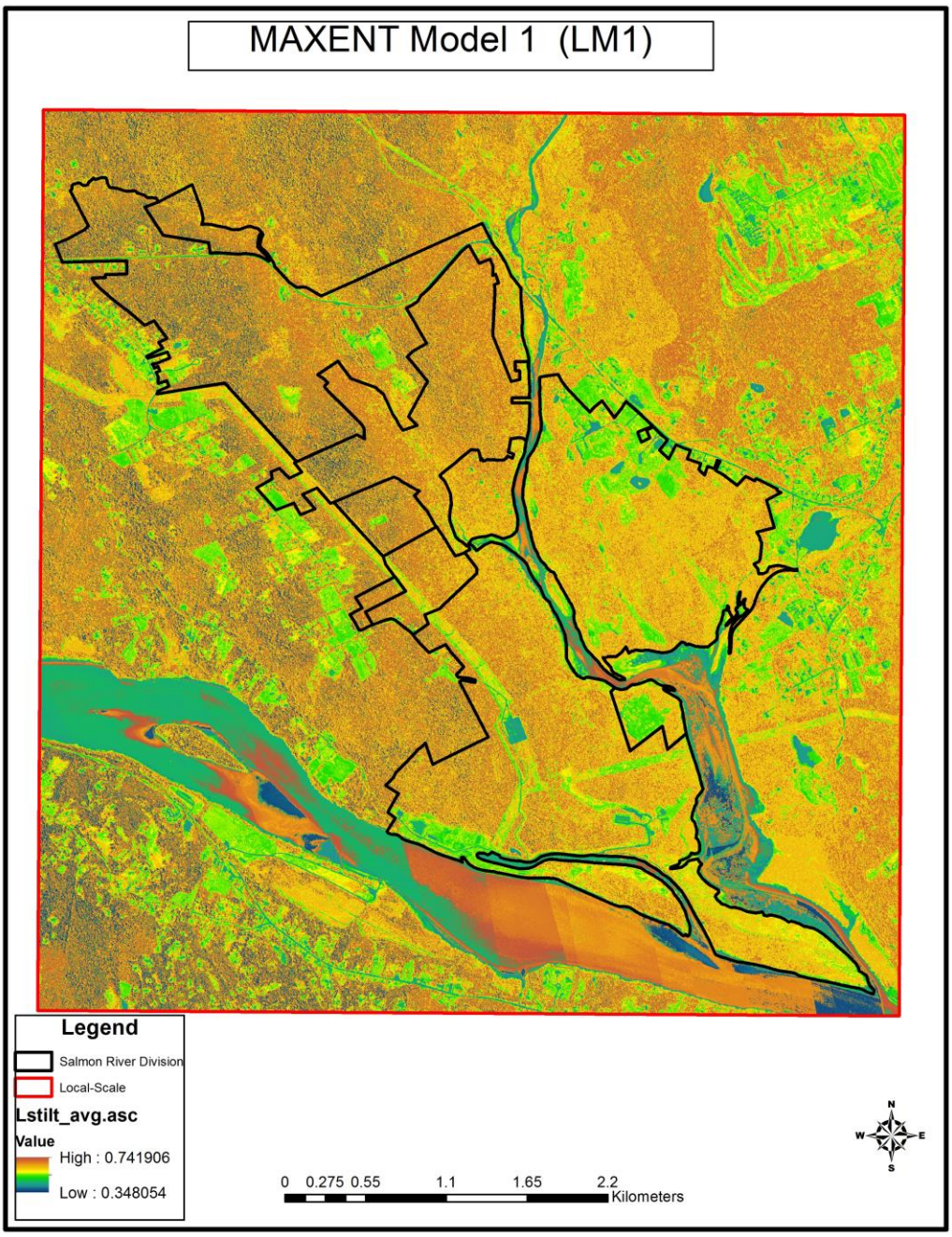


Figure 40: Map display of MaxEnt model 1 (LM1) results.



Evaluation

Each algorithm conducts internal validation measurements. The GLM produces an AIC score which is a measurement of the quality or goodness of fit of a statistical

model for a given set of data. Both ENFA and MaxEnt produce AUC scores which is a measurement of the true positive rate, or actual presence points, against the false-positive rate (1-true negative rate) or actual absence points – in this case pseudo-absence points. To understand how each algorithm perform against each other, model performance cannot simply be evaluated by the internal AIC or AUC score since the scoring values are set to different scales. To overcome this issue, model output values need to be set on a relative scale i.e., 0-1.

Approach

The results of each modeling algorithm's evaluation technique suggest that there is one model that performs best. The output raster values for each of the best performing models from each algorithm were rescaled to values ranging from 0 – 1. This rescaling allows for model evaluation to be performed among the algorithms. During the sample point clustering analysis, a total of 950 know presence points were left out of the modeling. Since there are known presence points and known pseudo-absence points, a confusion matrix was created to test the accuracy and precision of the model predictions that were correct. A threshold of 0.50 was applied to model output raster values. All values equal to or greater than the threshold value were considered "present", all values below the threshold were considered "absent".

- ***a*** is the number of **correct** predictions that an instance is **absent**
- ***b*** is the number of **incorrect** predictions that an instance is **present**
- ***c*** is the number of **incorrect** of predictions that an instance **absent**
- ***d*** is the number of **correct** predictions that an instance is **present**

Table 12: Example confusion matrix.

		Predicted	
		Absent	*Present*
Actual	*Absent*	***a***	***b***
	Present	***c***	***d***
	Accuracy = *Precision =*		

The *accuracy* is the proportion of the total number of predictions that were correct. It is determined using the equation:

The *precision* is the proportion of the positive predictions that are correct. It is determined using the equation:

Regional-Scale

At the regional-scale, the models that performed the best were GLM NEM5 with an AIC of 669.01, ENFA NEM5 with an AUC of 0.84. MaxEnt, however had two models that had the same AUC score of 0.912 (MAX NEM1 and MAX NEM4).

The confusion matrix (Table 13) depicts the best performing models for each algorithm and their respective prediction accuracy and precision. Although MaxEnt models NEM1 and NEM5 had the same AUC score of 0.912, MaxEnt NEM1 appears to be slightly better in prediction accuracy based off the confusion matrix. Although all model accuracy and precision was relatively good across all algorithms, MaxEnt model NEM1 had the highest predictive accuracy and precision of 0.800962 and 0.58 respectively, while the generalized linear model NEM5 predictive accuracy and precision was the lowest 0.759686 and 0.50 respectively. It appears that the GLM had a high false-

negative rate, or type II error, causing the model to under-estimate *M. vimineum's* distribution at the regional-scale.

Table 13: Regional-scale confusion matrix.

GLM NEM5			
		Predicted	
		Absent	*Present*
Actual	*Absent*	2975	25 TI(0.01)
	Present	924 TII(0.97)	25
	AC =	0.759686	
	PR =	0.50	

ENFA NEM5			
		Predicted	
		Absent	*Present*
Actual	*Absent*	2662	338 TI(0.11)
	Present	541 TII(0.57)	408
	AC =	0.777412	
	PR =	0.55	

MAXENT NEM1			
		Predicted	
		Absent	*Present*
Actual	*Absent*	2583	417 TI(0.14)
	Present	369 TII(0.39)	580
	AC =	0.800962	
	PR =	0.58	

MAXENT NEM4			
		Predicted	
		Absent	*Present*
Actual	*Absent*	2600	400 TI(0.13)
	Present	403 TII(0.42)	546
	AC =	0.796657	
	PR =	0.58	

Watershed-Scale

At the watershed-scale, the models that performed the best were GLM CTM4 with an AIC of 389.4836, ENFA CTM5 with an AUC of 0.60, and MAX CTM1 with an AUC of 0.935.

The confusion matrix (Table 14) depicts the best performing models for each algorithm and their respective prediction accuracy and precision. The GLM and ENFA model accuracy based off the confusion matrix scored relatively high (0.803 & 0.835 respectively) in comparison to the MaxEnt model who's accuracy was a mere 0.592. However, the ENFA had a higher false-positive rate, or type I error, causing the precision

to decrease. Due to a high false-positive rate (type II error), the MaxEnt model appears to over-estimate *M. vimineum* at the watershed scale. Since the ENFA has a higher accuracy rate than the GLM, but is less precise than the GLM, an accuracy paradox occurs.

The accuracy paradox states that a predictive model with a given level of accuracy may have greater predictive power than a model with a higher accuracy rate (Valverde-Albacete & Pelaez-Moreno, 2014). An example can be depicted with a dart board, where if player one has five darts and hits the target all five times, but the darts land in random locations within the target space (accuracy). Player two throws five darts, all darts land in a small portion of the target space, but not in the bull's eye (precision). To be accurate and precise, player three throws five darts, and all five darts land within the target's bull's eye. So, do we care about how often we hit the target (accuracy), or how often we hit the target's bull's eye (precision)?

Table 14: Watershed-scale confusion matrix.

		GLM CTM4	
		Predicted	
		Absent	*Present*
Actual	*Absent*	525	17 TI (0.03)
	Present	135 TII (0.58)	96
	AC =	0.803364	
	PR =	0.849	

		ENFA CTM5	
		Predicted	
		Absent	*Present*
Actual	*Absent*	490	52 TI(0.10)
	Present	75 TII(0.32)	156
	AC =	0.835705	
	PR =	0.750	

MAXENT CTM1			
		Predicted	
		Absent	*Present*
Actual	*Absent*	424	118 TI(0.22)
	Present	197 TII(0.85)	34
	AC =	0.592497	
	PR =	0.224	

Local-Scale

At the local-scale there was no single best performing model from each algorithm since NDVI was the only local-scale predictor value. Therefore, each algorithm's output was the "best" for each individual algorithm.

The confusion matrix (Table 15) depicts the best performing models for each algorithm and their respective prediction accuracy and precision. Although all accuracy test scores were fairly low, ENFA performed the best: *ac* =0.755. The GLM model had a high rate of false-positive instances, which indicate the model greatly over-estimated the presence of *M. vimineum* at the local-scale. The MaxEnt model displayed a very high rate of false-negative, or type II error, indicating the model under-estimated the distribution.

Table 15: Local-scale confusion matrix.

GLM LM1			
		Predicted	
		Absent	*Present*
Actual	*Absent*	220	780 TI(0.78)
	Present	8 TII(0.08)	87
	AC =	0.280365	
	PR =	0.100	

ENFA LM1			
		Predicted	
		Absent	*Present*
Actual	*Absent*	791	209 TI(0.21)
	Present	59 TII(0.62)	36
	AC =	0.755251	
	PR =	0.146	

MAXENT LM1			
		Predicted	
		Absent	*Present*
Actual	*Absent*	639	361 TI(0.36)
	Present	70 TII(0.73)	25
	AC =	0.606393	
	PR =	0.065	

Discussion and Conclusions

The modeling algorithms compared in this study, all of which exist as open-source platforms, may not be the most applicable modeling technique for predicting *M. vimineum's* occurrence distribution and potential suitable habitat. Although the algorithms employed in this study display varying degrees of performance at different scales, the algorithms rely on citizen-sourced presence-only occurrence data. These types of data are not collected according to a known sampling scheme. There is no randomized location selection process, or any consideration of spatial/temporal scales, and usually only conducted in convenient locations such as near roads and trails (Higby, Stafford, & Bertulli, 2012).

To better model *M. vimineum's* distribution and potential habitat, a more systematic randomized location sampling technique is needed that include true presence and true absence data and remove any sample selection bias (Phillips S. J., et al., 2009). The data sets used to compare the three models at three separate scales relies on a limited number of climate, landscape and topographic, and local predictor variables which, although discernably important for plant physiology and biology, may not be the most relevant for modeling *M. vimineum* distribution. Nor can there be a single or single group

of predictor variables that perfectly defines *M. vimineum's* distribution or its suitable habitat within the landscape. However, the large degree of variation captured with these results offers insight into the performance of these three modeling algorithms with presence-only information and scale-specific predictor variables.

Based on the results from comparing each modeling algorithm at each scale with respective predictor groups, there was a single "best" model for each scale. At the regional-scale, the MaxEnt algorithm had the highest relative accuracy and precision, followed by the ecological niche factor analysis algorithm and the generalized linear model. The ecological niche factor analysis algorithm displayed the highest accuracy at the watershed-scale, but the model's precision was lower than the generalized linear model's precision, thus leading to the accuracy paradox. Again, ecological niche factor analysis performed best at the local-scale with respect to accuracy and precision. However, the single predictor variable NDVI was not a significant variable in the prediction of *M. vimineum* within the generalized linear model. This was also confirmed in the AUC outputs from ENFA and MaxEnt. Obtaining more significant local-scale predictor variables may influence algorithm performance with respect to accuracy and precision; however, such fine-scale predictor variables are unlikely to be freely available. Nevertheless, these results support the alternative hypothesis that there exists a significant difference in modeling algorithm performance at different spatial scales with respect to accuracy and precision.

Owing to the fact that each modeling technique is open-source and relatively easy to implement, these results will likely be useful for land managers and conservation biologists with little statistical background to perform basic species distribution modeling

at several scales. The ability to model at different scales can be highly useful to anticipate the probability of a species', (in this case *M. vimineum),* habitat suitability and distribution with respect to land acquisition. Furthermore, understanding the probability of a species being present and or its probable habitat at several scales can influence where eradication efforts should take place to protect areas of high ecological diversity or known populations of rare, threaten, and or endangered species.

Results are also likely to suggest that there are scale-dependent strategies that can help reduce the spread of presently unknown populations of *M. vimineum.* For example, populations in Franklin county, MA. are currently the northern-most known range of *M. vimineum* within New England. Regional-scale results can be used to define a "battle front", or a distinct line on the landscape to understand where to deploy teams for eradication as *M. vimineum* marches northward.

Results based off the watershed-scale confusion matrix suggest that there are two competing models, the GLM model 4 (CTM4) with an accuracy = 0.80 and precision = 0.84, and ENFA model 5 (CTM5) with a slightly higher accuracy = 0.83 and yet lower precision of 0.75. Mentioning the accuracy paradox example, it would be more efficient for land managers to refer a model with a higher precision rate when deploying strike teams. Although the GLM is slightly less accurate, the precision rate is much higher. This suggests that land managers are more likely to deploy strike teams in areas that have a higher probability of occurrence and suitable habitat, thus increasing successful EDRR efforts. The watershed-scale ENFA model 5 results, based off the high specialization factor of 2.44, suggest that there may be hotspots of *M. vimineum* suitable habitat. These hotspots are useful to understand where dense clusters of suitable habitat might occur on

the landscape in relation to areas of high ecological integrity. For example, Dr. Kevin McGarigal and the Designing Sustainable Landscapes team in partnership with North Atlantic Landscape Conservation Cooperative have created an index of ecological integrity throughout the northeast region. Knowing that *M. vimineum* grows in a way forming dense patches inhibiting the regrowth of native species thus reducing biodiversity; we can overlay *M. vimineum* prediction to find out which areas of high ecological integrity are likely to be impacted. On the other hand, hotspots can also be useful to understand where areas of low density and less suitable habitat occur. These areas can be identified as target zones since less eradication effort is needed, thus reducing spread from already known locations of higher densities.

These scale-dependent strategies will help reduce eradication costs by identifying areas where eradication is likely to be successful given the models' results of the probability of *M. vimineum's* suitable habitat within the landscape. Results offer educational benefits, for example, the GLM model 4 (CTM4) offers insight to how elevation and distance to water features effects distribution. Rather than having a reactive management strategy, managers can employ a more active strategy to combat spread within areas that are known, based off model results, to offer suitable habitat.

This research explores *M. vimineum's* distribution at three distinct scales and sheds light on the factors which designate suitable habitat. However, further research involving inter-species transferability to understand model reaction would greatly increase EDRR efforts on other high-threat species. Comparing model accuracy and precision rates when results are extrapolated to a different geographic location given the current predictor variables would offer an understanding of *M vimineum's* distribution in

areas where predictor variable values are more extreme i.e., in areas where elevation and or temperature is more dramatic. Lastly, these results can be useful as a baseline for future prediction models applying climate change information. Applying known increases in temperature and other climate change information to the current predictor variables on a temporal scale would increase our understanding as to which factors effect spread over time with respect to climate change, thus enhancing future EDRR efforts

BIBLIOGRAPHY

Anderson, R. P., & Gonzalez Jr., I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling*, 2796-2811.

Berbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models, how, where, and hown many? *Methods in Ecology and Evolution*. doi:10.111/j.2041-210X-2011.00172.x

Blank, L., & Blaustien, L. (2012). Using ecological niche modeling to predict the distributions of two endangered amphibian species in aquatic breeding sites. *Hydrobiologia, 693*, 157-167. Retrieved 5 5, 2013

Boettner, C. (2013, April). (N. Bush, Interviewer) Sunderland, MA, United States.

Box, E. O. (1983). Macroclimate and plant Forms: An introduction to predictive modeling in phytogeography. *BioScience, 33*(6), 392. doi:10.2307/1309111

Bush, N. (2012). Personal Observations.

Chornesky, E. A., & Randall, J. M. (2003). The threat of Invasive Alien Species to Biological Diversity: Setting a Future Course. *Annals of the Missouri Botanical Garden, 90*(1), 67-76. Retrieved 1 7, 2014, from http://www.jstor.org/stable/3298527

Chornesky, E. A., & Randall, J. M. (2003). The Threat of Invasive Alien Species to Biological Diversity: Setting a Future Course. *Annals of the Missouri Botanical Garden, 90*(1), 67-76. Retrieved 12 3, 2013

Cole, P. G. (2003). *Environmental Constraints on the Distribution of the Non-native Grass,Microstegium vimineum.* PhD Dissertation, University of Tennessee. Retrieved 2 3, 2013, from http://trace.tennessee.edu/utk_graddiss/1494

Connecticut River Invasive Species Partnership. (2014). *Identifying priority areas for invasive plant management in the Connecticut River watershed.*

Council for Agriculture Science and Technology. (2002, March). Invasive Pest Species: Impascts on Agriculture Production, Natural Resources,and the Environment. *Issue Paper*(20). Retrieved from http://www.cast-science.org/publications/?invasive_pest_species_impacts_on_agricultural_production_natural_resources_and_the_environment&show=product&productID=2863

Ehrenfeld, J. G. (1999). A rhizomatous, perennial form of Microstegium vimineum (Trin.) A. Camus in New Jersey. *Journal of the Torrey Botanical Society, 126*(4), 352-358. Retrieved 2 3, 2013

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 677-697. doi:10.1146/annurev.ecolsys.110308.120159

Elith, J., Phillips, S. J., Hastie, T., Dudlik, m., Chee, Y., & Yates, C. J. (2011). A staistical explanation of MaxEnt for ecologists. *Diversity and Distributions, 17*, 43-57. Retrieved 5 1, 2013

Engler, R., Guisan, A., & Rechstiener, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology, 41*, 263–274. Retrieved 1 27, 2014, from http://onlinelibrary.wiley.com.silk.library.umass.edu/doi/10.1111/j.0021-8901.2004.00881.x/pdf

Erfanian, B., Hamed Mirkarimi, S., Salman Mahini, A., & Reza Rezaei, H. (2013). A presence-only habitat suitability model for Persian leopard Panthera pardus saxicolor in Golestan National Park, Iran. *Wildlife Biology, 19*(2), 170-178. doi:http://dx.doi.org.silk.library.umass.edu/10.2981/12-045

ESRI. (2013). *Arc GIS for desktop*. Retrieved from ESRI Products: http://www.esri.com/software/arcgis/arcgis-for-desktop

Gibson, D. J., & Benedict, J. (2002). Life history of Microstegium vimineum (Poaceae), an invasive grass in southern Illinois. *Journal of the Torrey Botanical Society, 129*(3), 207-219. Retrieved 2 3, 2013

Gotelli, N. J., & Ellison, A. M. (2013). *A Primer of Ecological Statistics* (Second Edition ed.). Sunderland, MA: Sinauer Associates, Inc.

Higby, L. K., Stafford, R., & Bertulli, C. G. (2012). An Evaluation of Ad Hoc Presence-Only Data in Explaining Patterns of Distribution: Cetacean Sightings from Whale-Watching Vessels. *International Journal of Zoology, 2012*, 1-5. doi:10.1155/2012/428752

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological Niche Factor Analysis: How to Compute Habitat-Suitablity Mpas without Absence Data? *Ecology, 83*(7), 2027-2036. Retrieved 1 15, 2013

Hirzel, A., Hausser, J., & Perrin, N. (2004). Biomapper 3.1. Lab. of Conservation Biology, Department of Ecology and Evolution, University of Lausanne. Retrieved 1 28, 2014, from Biomapper: A GIS-toolkit to model ecological niche and habitat suitability: http://www2.unil.ch/biomapper/

Hunt, D. M., & Zaremba, R. E. (1992). The northeastward spread of Microstegium vimineum (Poaceae) into New York and adjacent states. *Rhodora, 94*(878), 167-170. Retrieved 12 1, 2013

Ibanez, I., Silander, J. A., Allen, M., Treanor, S. A., & Wilson, A. (2009). Identifying hotspots for plant invasions and forcasting focal points of further spread. *Journal of Applied Ecology, 46,* 1219-1228. Retrieved 2 3, 2013

Invasive Plant Control Inc. (2011, 7 25). *IPCGSAAdvantageTextFinal.pdf.* Retrieved 1 16, 2014, from invasiveplantcontrol.com: http://www.invasiveplantcontrol.com/pdflinks/IPCGSAAdvantageTextFinal.pdf

Kumar, S., & Stohlgren, T. J. (2009). Maxent modeling for predicting suitable habitat for threatened and endangered tree Canacomyrica monticola in New Caledonia. *Journal of Ecology and Natural Environment, 1*(4), 094-098. Retrieved 10 4, 2013

Mack, R., Simberloof, D., Lonsdale, W., Evans, H., Cout, M., & Bazzaz, F. (2000). Biotic invasions: Causes, epidemiology, global consequences, and control. *Ecological Applications*, 689-710.

McGarigal, K. (2013). BioStats. Massachusetts, United States.

Mehrhoff, L. (2000). Immigration and expansion of the New England flora. *Rhodora, 102*, 280-298. Retrieved 1 1, 2013

Moisen, G. G., & Frescino, T. S. (2002). Comparing Five Modelling Techniques for Predicting Forest Characteristics. *Ecological Modelling*, 209-225. Retrieved 1 20, 2014

Morin, X., & Thuiller, W. (2009). Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. *Ecology, 90*(5), 1301–1313. Retrieved 2 3, 2014

Munoz, M. E., Giovanni, R., F, S. M., Sutton, T., Pereira, R. S., Canhos, D. A., & Canhos, V. P. (2011). OpenModeller: A Generic Approach to Species' Potential Distribution Modeling. *GeoInformatica, 1*, 111-135. doi: 10.1007/s10707-009-0090-7

National Wildlife Refuge System. (2003). *The national strategy for managemnet of invasive species.* National Invasive Species Management Strategy Team. Retrieved 2 3, 2014, from http://www.fws.gov/invasives/pdfs/NationalStrategyFinalRevised05-04.pdf

Neeti, N., Vaclavik, T., & Niphadkar, M. (2007, June). Potential Distribution of Japanese KNotweed (Polygonum cuspidatum) in Massachusetts.

New York Department of Environmental conservation . (2013). *Health Hazards & Safety Instructions for Giant Hogweed*. Retrieved 1 16, 2014, from NY Department of Environmental Conservation : http://www.dec.ny.gov/animals/72556.html

NJ Invasive Species Strike Team. (2014). *New Jersey invasive species strike team*. Retrieved from New Jersey invasive species strike team: http://www.njisst.org/index.asp

Northwest Alliance for Computational Science and Engineering. (2015). *PRISM CLimate Group*. Retrieved from PRISM Climate Data: http://www.prism.oregonstate.edu/

Office of Geographic information. (2015). *The Offical Website of the Executive Office for Administration and Finance*. Retrieved from http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/

Ortega-Huerta, M. A., & Townsend Peterson, A. (2008). Modeling Ecological Niches and Predicting Geographical Distributions: A Test of Six Presence-only Methods. *Revista Mexicana de Biodiversidad*, 205-216.

Paini, D. R., Bianchi, F. J., Northfield, T. D., & De Barro, P. J. (2011). Predicting Invasive Fungal Pathogens Using Invasive Pest Assemblages: Testing Model Predictions in a Virtual World. *PLoS One, 6*(10). Retrieved 1 20, 2014

Pajchar, L., & Mooney, H. A. (2009). Invasive Species, Ecosystem Services and Human Well-being. *Trends in Ecology and Evolution, 24*(9), 497-504. Retrieved 1 7, 2014, from http://www.environment.ucla.edu/media_IOE/files/Pejchar-and-Mooney-2009---invasives-and-ecosystem-services-ls-y5k.pdf

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*, 231–259. Retrieved 1 27, 2014, from http://www.cs.princeton.edu/~schapire/papers/ecolmod.pdf

Phillips, S. J., Dudik, M., & Schapire, R. E. (2004). A Maximum Entropy Approach to Species Distribution Modeling. *Twenty-First International Conference on Machine Learning*, (pp. 655-662).

Phillips, S. J., Dudik, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data. *Ecological Applications, 19*(1), 181-197.

Pimental, D., Lach, L., Zungia, R., & Morrison, D. (2000). Environmental and economic costs of nonindigenous species in the United States. *BioScience*, 53-65.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the Environmental and Economic Costs Associated with Alien-Invasive Species in the United States. *Ecological Economics, 52*, 273-288. Retrieved 1 9, 2014

Quinn, G. P., & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists.* New ork: Cambridge University Press.

R Core Team . (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Rupprecht, F., Oldeland, J., & Finckh, M. (2011). Modelling Potential Distribution of the Threatened Tree Speices Juniperus oxycedrus: How to Evaluate the Predictions of Different Modelling Approaches? *Journal of Vegetation Science, 22*, 647-659. Retrieved 1 20, 2014

Silvio O. Conte National Fish and Wildlife Refuge. (2014). *Silvio O. Conte National Fish and Wildlife Refuge: Invasive Species*. Retrieved from Silvio O. Conte National Fish and Wildlife Refuge: Northeast Region: http://www.fws.gov/r5soc/invasive_species/index.html

StatSoft. (2014). *How to Reduce Number of Variables and Detect Relationships, Principal Components and Factor Analysis*. Retrieved 1 28, 2014, from StatSoft: http://www.statsoft.com/Textbook/Principal-Components-Factor-Analysis

Syphard, A. D., & Franklin, J. (2009). Difference in Spatial Predictions Among Species Distribution Modeling Mehtods Vary with Species Traits and Environmental Predictors. *Ecography, 32*, 907-918. Retrieved 2 1, 2013

Tennessee Exotic Pest Plant Council . (2013, 2 3). *Invasive Plants*. Retrieved from Tennessee Exotic Pest Plant Council: http://www.tneppc.org/invasive_plants/58

Thuiller, W., Lafourcade, B., Engler, R., & Araujo, M. B. (2009). BIOMOD: A platform for esemble forcasting of species distributions. *Ecography, 32*, 369-373. Retrieved 10 2, 2013

U.S. Congress, Office of Technology Assessment. (1993). *Harmful Non-Indigeous Species in the United States, OTA-F-565*. Washington DC. : U.S. Government Printing Office.

U.S. National Invasive Species Council (NISC). (2013). *Fiscal Year 2012 Invasive Species Interagency Crosscut Budget*. Retrieved 1 15, 2014, from http://www.invasivespecies.gov/global/org_collab_budget/org_collab_budget_documents/NISC_2012_Crosscut_Budget_Summary.pdf

United Sates Fish and Wildlife Service. (2012, 1). *The cost of invasive species* . Retrieved from U.S. Fish & Wildlife Service.

United States Department of Agriculture. (2012, April). Giant Hogweed (Heracleum mantegazzianum). USDA. Retrieved from Ahpis.

United States Department of Agriculture, Farm Service Agency. (2015). *NAIP Imagery*. Retrieved from http://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/index

United States Department of Agriculture, Forest Service. (2004). *National strategy and implementation plan for invasive species management.* USDA. Retrieved 2 3, 2014, from http://www.fs.fed.us/invasivespecies/documents/Final_National_Strategy_100804.pdf

United States Fish and Wildlife Service. (2009, 2 4). *What is an invasive species*. Retrieved from Invasive Species: http://www.fws.gov/invasives/

United States Geological Survey. (2004). *Invasive species program: Five year program plan.* USGS. Retrieved 2 3, 2014, from http://www.usgs.gov/ecosystems/invasive_species/USGSInvasiveSpeciesProgramFiveYearProgramPlanFiscalYears2005-2009.pdf

United States Geological Survey. (2014, January 13). *Hydrography*. Retrieved from The National Map: http://nhd.usgs.gov/

University of Georgia, Center for Invasive Species and Ecosystem Health. (2014, 1 27). *EDDMapS, Early Detection and Distribution Mapping System*. Retrieved 1 27, 2014, from EDDMapS.org: http://www.eddmaps.org/

University of Georgia, Center for Invasive Species and Ecosystem Health. (2014, 1 27). *IPANE, Invasive Plant Atlas of New England*. Retrieved 1 27, 2014, from IPANE: http://www.eddmaps.org/ipane/

University of Massachusetts. (2000). *UMass landscape Ecology Lab*. Retrieved from Designing Sustainable Landscapes: http://www.umass.edu/landeco/research/dsl/dsl.html

University of Massachusetts. (2015). *MassWoods Forest Conservation Program*. Retrieved from The Outsmart Invasives Species Project: http://masswoods.net/outsmart

USDA, F. S. (2015, 6 20). *Microstegium vimineum*. Retrieved from USFS database: http://www.fs.fed.us/database/feis/plants/graminoid/micvim/all.html#201

Vaclavik, T., & Ortega, M. (2008, April). Modeling potential distributionof Norway maple (Acer platanoides) in Massachusetts, USA.

Valverde-Albacete, F., & Pelaez-Moreno, C. (2014). 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factore Explains the Accuracy Paradox. *PLOS one*. doi:10.1371/journal.pone.0084217

Vitousek, P. M. (1990, February). Biological invasions and ecosystem processes: towards an integration of population biology and ecosytem studies. *Oikos, 57*(1), 7-13. Retrieved 1 14, 2014, from http://www.jstor.org/stable/3565731.

Vitousek, P. M., D'Antonio, C. M., Loope, L. L., & Westbrooks, R. (1996). Biological Invasions as Global Environmental Change . *American Scientists, 84*(5), 468-478. Retrieved 1 15, 2014, from http://www.jstor.org.silk.library.umass.edu/stable/pdfplus/29775751.pdf?acceptTC=true&acceptTC=true&jpdConfirm=true

Wang, M., Wright, J., Buswell, R., & Brownlee, A. (2013). A COMPARISON OF APPROACHES TO STEPWISE REGRESSION FOR GLOBAL SENSITIVITY ANALYSIS USED WITH EVOLUTIONARY OPTIMIZATION. *13th Conference of International Building Performance Simulation Association*, (pp. 2551-2558). Chambéry. Retrieved 1 28, 2014, from http://www.ibpsa.org/proceedings/BS2013/p_1047.pdf

Weaver, J. E., Conway, T. M., & Fortin, M. (2012). An invasive species' relationship with environmental variables changes across multiple spatial scales. *Lanscape Ecology, 27*(9), 1351-1362. Retrieved 1 23, 2014, from http://link.springer.com.silk.library.umass.edu/article/10.1007/s10980-012-9786-4

Wilcove, D. S., Rothstein, D., Dubow, J., Phillips, A., & Losos, E. (1998). Quantifying Threats to Imperiled Spceies in the United States. *Bioscience, 48*(8), 607-615. Retrieved 1 9, 2014

Wilcove, D. S., Rothstein, D., Dubow, J., Phillips, A., & Losos, E. (1998). Quantifying threats to imperiled species in the United States. *BiScience, 48*(8), 607-615. Retrieved 1 2, 2014

Wolfe, B. E., Rodgers, V. L., Stinson, K. A., & Pringle, A. (2008). The invasive plant Alliaria petiolata (garlic mustard) inhibits ectomycorrhizal fungi in its introduced range. *Journal of Ecology, 96*, 777-783. doi:10.1111/j.1365-2745.2008.01389.x

Xuezhi, W.; Weihua, X.; Zhiyun, O; Jianguo, L.; Yi, X.; Youping, C. (2008). *Acta Ecologica Sinica,, 28*(2), 821-828. Retrieved 5 23, 2013